STATISTICAL METHODS FOR HIGH DIMENSIONAL BIOMEDICAL DATA

A Dissertation

by

ROBYN LYNN BALL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Alan R. Dabney
Committee Members,	Raymond J. Carroll
	Alan H. Feiveson
	Susan C. Geller
Department Head,	Simon Sheather

May 2013

Major Subject: Statistics

Copyright 2013 Robyn Lynn Ball

ABSTRACT

This dissertation consists of four different topics in the areas of proteomics, genomics, and cardiology. First, a data-based method was developed to assign the subcellular localization of proteins. We applied the method to data on the bacteria *Rhodobacter sphaeroides* 2.4.1 and compared the results to PSORTb v.3.0. We found that the method compares well to PSORTb and a simulation study revealed that the method is sound and produces accurate results. Next, we investigated genomic features involved in the lethality of the knockout mouse using the random forest technique. We achieved an accuracy rate of 0.725 and found that among other features, the evolutionary age of the gene was a good predictor of lethality. Third, we analyzed DNA breakpoints across eight different cancer types to determine if common hotspots or cancer-type specific hotspots can be well-predicted by various genomic features and investigated which of the genomic features best predict the number of breakpoints. Using the random forest technique, we found that cancer-type specific hotspots are poorly predicted by genomic features but common hotspots can be predicted using the relevant genomic features. Additionally, we found that among the genomic features analyzed, indel rate and substitution rate were consistently chosen as the top predictors of breakpoint frequency. Lastly, we developed a method to predict the hypothetical heart age of a subject based on the subject's electrocardiogram (ECG). The heart age predictions are consistent with current ECG science and knowledge of cardiac health.

For Justin Ball, who has always seen me as perfect and for Elaine Phares, who knows I am not.

ACKNOWLEDGEMENTS

First, I wish to thank Dr. Alan Dabney for his consistent support. He has been kind, always helpful, and generous with his expertise, time, and energy. I also wish to thank Dr. Raymond Carroll and Dr. Al Feiveson for all the help they have given me throughout this process. Dr. Carroll has always been willing to help, even though I know he is extremely busy. He has given me wonderful advice throughout my graduate studies. Dr. Feiveson has been a joy to work with at NASA and he has taught me much about statistics. Additionally, I wish to thank Dr. Sue Geller for serving on my committee and for her timely advice. In addition to my chair and committee, I want to thank Dr. Michael Longnecker. Dr. Longnecker has provided me with wonderful guidance and has always joyfully supported me in my endeavors. He is a true gift to this department.

Many others have given their expertise to the projects in this dissertation. Dr. Stephen Callister has been an invaluable resource for the proteomics project. Dr. Han Liang at UT MD Anderson provided me with interesting projects that later turned into published papers. He has been very supportive and it was wonderful to work with him, Yuan, Yudong, Jun, and Leng. At NASA, Dr. Todd Schlegel provided the project his extensive expertise in cardiology and electrocardiography. He has been a joy to work with and I appreciate his continued support.

My family and friends have helped me immensely. Thank you Mom, Bob, Chris, Lisa, Lexi, Emily, Dad, and Julie. I particularly want to thank Rob, Suzanne, and Mabeen for seeing me through the rough times, Lori for her encouragement, and all of my other friends for cajoling me through this process.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

My dissertation involved four topics in the areas of proteomics, genomics, and cardiology:

1. A data-based method for assigning the subcellular localization of proteins,
2. Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data,
3. Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots,
4. Predicting "heart age" using the electrocardiogram (ECG).

I address each topic, in turn, in sections two through five. In section two, I present a method that assigns the subcellular localization to proteins based on mass spectrometry (MS) data. The third section discusses the predictability of lethality in the knockout mouse using genomic features and the fourth section investigates genomic features involved in cancer hotspots. In section five, I present a method to predict a subject's "heart age" (hypothetical age) based on the electrocardiogram (ECG).

2. A DATA-BASED METHOD FOR PROTEIN SUBCELLULAR LOCALIZATION

2.1 Introduction

Information regarding the subcellular localization of a protein is of significant interest to scientists studying microorganisms. A subcellular fraction is a compartment within a cell, and proteins may be present in a single fraction or present in multiple fractions. Empirical observations of localized proteins, based upon mass spectrometry (MS) analysis of subcellular fractions, often result in proteins being observed across several fractions. The reasoning for this observation is a lack of a pure subcellular fraction and that proteins in relatively high abundance can contaminate multiple subcellular fractions. Scientists need an automated statistical procedure to test for protein presence based upon empirical evidence. Current methods often necessitate some prior knowledge of the protein or may only be applicable to highly abundant proteins (Callister et al., 2006a). In addition, many methods are tailored for a specific organism, are only for gram positive or gram negative bacteria, or only identify proteins that are in a single fraction. We developed a data-based method that is free of these restrictions. It can be applied to any organism for which these type of data are available, and it identifies proteins in single or multiple fractions with statistical confidence.

Our method combines peptide peak intensity information as a measure of protein presence in a subcellular fraction. For each fraction, the relative abundance of the protein is estimated as well as its standard error. From these estimates, a test statistic and p-value are computed that allow us to decide whether the protein is present in the fraction. If a protein is present in more than one fraction, Welch's t-test or Hsu's

multiple comparison procedure is employed to estimate the fraction, or fractions, in which the protein is most abundant. This allows scientists to determine not only if a protein is present but, more specifically, the fraction(s) in which the protein most likely resides.

We applied this method to data collected on the bacteria *R. sphaeroides* 2.4.1 and compared the assignments to predictions from PSORTb v.3.0 (Yu et al., 2010). PSORTb makes final predictions by combining the predictions of six modules of varying methodology. While PSORTb's predictions are based on amino acid sequences, our method, in contrast, makes assignments based on observed MS data, which should always be more precise. To further evaluate our method, we performed a simulation study.

2.2 Methods

2.2.1 Experimental methods

All experimental methods were carried out at the Pacific Northwest National Laboratory (PNNL). Scientists at PNNL stated that culture conditions for *R. sphaeroides* 2.4.1 chemoorganoheterotrophic (non-photosynthetic) and photoorganoheterotrophic (photosynthetic) growth as well as protocols for the fractionation of harvested cells into subcellular fractions (cytoplasm, periplasm, cytoplasmic membrane, outer membrane, and intracytoplasmic membrane) have been previously published (Barber et al., 1996; Cohen-Bazire et al., 1957; Deal and Kaplan, 1983; Flory and Donohue, 1995; Jackson, 1991; Tai and Kaplan, 1985; Weiss, 1976). The extraction of proteins from each subcellular fraction followed protocols previously established for conducting proteomics analyses utilizing the bottom-up, or peptide centric approach (Callister et al., 2006b).

2.2.2 Statistical methods

One of the challenges of using MS data is that there are frequently many zeros due to absent or low abundant peptides that are not detected by the instrument. Instead of throwing out the zeros, we designed a statistical model that accounts for the presence of zeros. For each fraction, twelve samples were taken in the non-photosynthetic condition and fifteen samples were taken in the photosynthetic condition. Each sample had three technical replicates. Statistical Tools for AMT Tag Confidence (STAC) is a method that quantifies the confidence in the AMT tag identification, where AMT stands for Accurate Mass and Time (Stanley et al., 2011). Based on expert advice from the biologists at PNNL, only data on peptides that had a mass tag STAC Score ≥ 0.90 and a fully-tryptic cleavage were used in this study. The abundance, Y, of the peptide in a sample and fraction was taken to be the natural logarithm of the average of the observed unscaled mass tag peak intensities across technical replicates.

The method is a two-step method. In step one, we determine in which fractions the protein is present and in step two, we make the final assignment of where the protein is most abundant. We first identify the peptides that were observed in each fraction that also map to the protein. For each of these peptides, we then estimate the peptide-level mean abundance as the average of observed peptide abundances over the sample. For example, if the peptide was observed in 10 of the 12 samples, the peptide-level mean abundance for this protein is the sum of the observed abundances divided by 12. Next, we estimate the protein abundance and the standard error of the protein abundance in each fraction. The estimated protein abundance in the fraction is the simple average of the estimated peptide-level mean abundances that map to that protein. The next section describes the full mathematical details of our method, including the details for estimating the standard error of protein abundance

estimates.

In step one, we determine if a protein is present in a fraction by calculating the t-statistic that would correspond to testing the null hypothesis that the mean protein abundance is zero in the fraction versus the one-sided alternative hypothesis, that the mean protein abundance is greater than zero in the fraction. Practically speaking, the mean protein abundance cannot be zero since we are working with observed data but the t-statistic still provides a useful index for deciding whether or not the protein is present in the fraction. We decided that the protein was present in the fraction if the p-value was less than $\alpha = 0.01$.

Step two assigns the protein to the fraction(s) in which the protein is most abundant. If the protein is present in only one fraction, we assign the protein to that fraction. However, consider the case for which the protein is present in multiple fractions. We would like to make a call as to where the protein is most abundant and the likely subcellular fraction of primary localization. If the protein is present in two fractions, we perform Welch's t-test for unequal sample sizes and unequal variances (Welch, 1947) to decide if the protein abundance for one fraction is higher than the protein abundance for the other fraction. If it is, we assign the protein to the fraction with the highest abundance. If there is not a significant difference in the abundances, we assign the protein to both fractions. If the protein is present in more than two fractions, we use Hsu's procedure for multiple comparisons with unequal sample sizes (Hsu, 2006) at the $\alpha = 0.01$ level to determine the fraction, or fractions, in which the protein is most abundant. The mathematical details for carrying out this procedure are given next.

2.2.2.1 Statistical method details

Let x_{ijkst} be the peptide peak intensity for protein i, peptide j, fraction k, sample s, and technical replicate t. Thus, the peptide abundance is $Y_{ijks} = \log(\frac{1}{m}\sum_{t=1}^{m} x_{ijkst})$ for protein i, peptide j, fraction k, and sample s where $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, m_{ik}$, $k = 1, 2, \ldots, K$, $s = 1, 2, \ldots, n$, and $t = 1, 2, \ldots, m$. We set $Z_{ijks} = 1$ if the peptide is observed in the sample and $Z_{ijks} = 0$ if the peptide is not observed. In reality, it is possible that $Z_{ijks} = 1$ but the peptide is mis-identified or that $Z_{ijks} = 0$ but the peptide is actually present; it is possible that we have left censoring in that low abundances may not be observed. In the simulation study, we evaluated the method under these conditions.

Let μ_{ijk} be the mean abundance of peptide j corresponding to protein i in fraction k over all hypothetical samples. Using properties of conditional expectation, we can write:

$$\begin{aligned}
\mu_{ijk} &= \mathrm{E}(Y_{ijks}) \\
&= \mathrm{E}(Y_{ijks}|Z_{ijks} = 1) \times \Pr(Z_{ijks} = 1) + \mathrm{E}(Y_{ijks}|Z_{ijks} = 0) \times \Pr(Z_{ijks} = 0) \\
&= \theta_{ijk} \times \tau_{ijk} + \epsilon_{ijk}
\end{aligned}$$

where ϵ_{ijk} represents the expected abundance of the peptide in the case of censoring, $\tau_{ijk} = \Pr(Z_{ijks} = 1)$ is the probability that peptide j that corresponds to protein i is present in a randomly-selected sample from fraction k, and $\theta_{ijk} = \mathrm{E}(Y_{ijks}|Z_{ijks} = 1)$ is the expected abundance for this feature, given that it is present in the sample.

A natural estimate for θ_{ijk} is the average of the observed abundances of peptide

j corresponding to protein i in fraction k:

$$\hat{\theta}_{ijk} = \frac{\sum_{s=1}^{n} Z_{ijks} Y_{ijks}}{\sum_{s=1}^{n} Z_{ijks}} = \frac{1}{n_{ijk}} \sum_{s=1}^{n} Z_{ijks} Y_{ijks},$$

where $n_{ijk} = \sum_{s=1}^{n} Z_{ijks}$ is the number of times this feature was observed in the samples. A natural estimate for τ_{ijk} is the proportion of times peptide j corresponding to protein i was identified in fraction k: $\hat{\tau}_{ijk} = n_{ijk}/n$. Thus, we can estimate μ_{ijk} as:

$$\hat{\mu}_{ijk} = \hat{\theta}_{ijk} \times \hat{\tau}_{ijk} = \frac{1}{n} \sum_{s=1}^{n} Z_{ijks} Y_{ijks}.$$

Note that this differs from the simple average of the observed intensities, which would equal $(n/n_{ijk}) \times \hat{\mu}_{ijk}$. In cases where a peptide is observed in some samples but not others for a given fraction, $\hat{\mu}_{ijk}$ will attenuate the simple average toward zero, taking into account the frequency of unobserved intensities.

Let ω_{ik} be the protein-level mean abundance for protein i in fraction k. This can be expressed as the average of the peptide-level expectations:

$$\omega_{ik} = \frac{1}{m_{ik}} \sum_{j=1}^{m_{ik}} \mu_{ijk}.$$

We estimate ω_{ik} as:

$$\hat{\omega}_{ik} = \frac{1}{m_{ik}} \sum_{j=1}^{m_{ik}} \hat{\mu}_{ijk}.$$

To derive the standard error of $\hat{\mu}_{ijk}$, we compute:

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{ijk}) &= \frac{n}{n^2}\Big\{\mathrm{E}_Z\Big[\mathrm{Var}\big(Z_{ijks}Y_{ijks}|Z_{ijks}=z_{ijks}\big)\Big] \\
&\qquad +\mathrm{Var}_Z\Big[\mathrm{E}\big(Z_{ijks}Y_{ijks}|Z_{ijks}=z_{ijks}\big)\Big]\Big\} \\
&= \frac{1}{n}\Big\{\mathrm{E}_Z\Big[(1-Z_{ijks})\times 0 + Z_{ijks}\times\sigma^2_{ijk}\Big] \\
&\qquad +\mathrm{Var}_Z\Big[(1-Z_{ijks})\times 0 + Z_{ijks}\times\theta_{ijk}\Big]\Big\} \\
&= \frac{1}{n}\Big\{\sigma^2_{ik}\times\tau_{ijk} + \theta^2_{ijk}\times\tau_{ijk}(1-\tau_{ijk})\Big\}
\end{aligned}$$

where

$$\sigma^2_{ijk} = \mathrm{Var}(Y_{ijks}|Z_{ijks}=1)$$

which is estimated by

$$\hat{\sigma}^2_{ijk} = \frac{1}{n_{ijk}-1}\sum_{s=1}^{n}(Y_{ijks}Z_{ijks}-\hat{\theta}_{ijk})^2 Z_{ijks}.$$

We can estimate σ^2_{ik} as the pooled sample variance of the observed abundances:

$$\hat{\sigma}^2_{ik} = \frac{\sum_{j=1}^{m_{ik}}(n_{ijk}-1)\hat{\sigma}^2_{ijk}}{\sum_{j=1}^{m_{ik}}(n_{ijk}-1)}.$$

We estimate $\mathrm{Var}(\hat{\mu}_{ijk})$ by substituting the estimates for μ_{ijk}, σ_{ij}, and τ_{ijk}. Similarly, the variance of the protein-level mean abundance is:

$$\mathrm{Var}(\hat{\omega}_{ik}) = \frac{1}{m^2_{ik}}\sum_{j=1}^{m_{ik}}\mathrm{Var}(\hat{\mu}_{ijk})$$

and can be estimated by substituting the estimates $\hat{\mathrm{Var}}(\hat{\mu}_{ijk})$.

For the purposes of deciding whether protein i is present in the population at fraction k, we calculate the t-statistic

$$t_{ik} = \frac{\hat{\omega}_{ik}}{\hat{s.e.}(\hat{\omega}_{ik})},$$

with

$$df_{ik} = \left(\sum_{j=1}^{m_{ik}} n_{ijk} \right) - m_{ik}$$

where $\hat{\omega}_{ik}$ and $\hat{s.e.}(\hat{\omega}_{ik})$ are estimates of the parameter and its standard error, respectively. We then assign p-values

$$p_{ik} = 1 - \Pr(t_{ik} < T),$$

where $T \sim t_{df_{ik}}$. We decide that there is significant evidence that protein i is present in fraction k if $p_{ik} \leq \alpha_1$. For this study, we set $\alpha_1 = 0.01$ to be conservative but one could increase α to have a greater chance of identifying subcellular localizations.

Now we perform Step 2. In some cases, $p_{ik} \leq \alpha_1$ in multiple fractions so a decision must be made to determine the set of fractions, G_i, where the protein is most abundant. Choose α_2 at which to do the following tests. We chose $\alpha_2 = \alpha_1 = 0.01$. For protein i, let

$$\hat{G}_i = \{k : \hat{\omega}_{ik} \text{ are not significantly different}\}$$

If the protein is present in two fractions at the α_1 level, use Welch's t-test (Welch, 1947) for unequal sample sizes and unequal variance to compare the estimated protein abundances. If the resulting p-value $\leq \alpha_2$ we decide that the fraction corresponding to the highest abundance is in $\hat{G}_i$. Otherwise, both fractions are in $\hat{G}_i$. If the protein

is present in more than two fractions, use Hsu's multiple comparison procedure for unequal replications at the α_2 level (Hsu, 2006). Based on these results, include or exclude locations in $\hat{G}_i$.

When performing Hsu's procedure (Hsu, 2006), let $H_i = \{k : p_{ik} \leq \alpha_1\}$. Instead of using the MSE, let the pooled sample variance be calculated as follows:

$$\hat{\sigma}^2 = \frac{\sum_{k \in H_i} df_{ik} \hat{\sigma}^2_{ik}}{\sum_{k \in H_i} df_{ik}},$$

so that the total degrees of freedom for error is $\nu = \sum_{k \in H_i} df_{ik}$. This is appropriate, in this case, since we estimate ω_{ik} by estimating m_{ik} μ_{ijk}'s and not merely by the sample average. By doing so, we account for the loss in the degrees of freedom.

Let $t = |H_i|$, $c = 1, 2, \ldots, t$ and $s = 1, 2, \ldots, t-1 : s \neq c$. For $c = 1, 2, \ldots, t$ and $k \in H_i$, let $\hat{\omega}_{ik} = \hat{\mu}_c$ so that the cth location is the control for Dunnett's multiple comparison with a control (MCC) procedure. Compute:

$$\hat{\mu}_c - \hat{\mu}_s + d^c \hat{\sigma} \sqrt{\frac{1}{n_{ic}} + \frac{1}{n_{is}}}$$

$$d^c = m d_{\alpha_2, t-1, \nu}$$

such that $d_{\alpha_2, t-1, \nu}$ is Dunnett's one-sided critical value at level α_2 and

$$m = \max_{s \neq c} \{1 + 1.07(1 - \frac{n_{is}}{n_{ic}})\}$$

. Once this is completed for all $s \neq c$, let

$$D_c^+ = \min_{s \neq c} \left\{ \hat{\mu}_s - \hat{\mu}_s + d^c \hat{\sigma} \sqrt{\frac{1}{n_{ic}} + \frac{1}{n_{is}}} \right\}$$

. If $D_c^+ < 0$ set $D_c^+ = 0$. Repeat this for each of $c = 1, 2, \ldots, t$ so that we have t

upper bounds equal to D_c^+. If $D_c^+ > 0, k \in G_i$.

2.2.3 Simulation study

A simulation study was carried out to test the validity of the two-step process of assigning protein localization. First, the simulated data were generated such that there were 100 proteins assigned to each subcellular fraction. The first 100 proteins were assigned to the periplasm, the second 100 proteins were assigned to the cytoplasm, and so on. Call the fraction where the protein was assigned the "target fraction" and all other fractions where the protein was not assigned the "non-target" fractions where we simulate false positives in the data. In addition to false positives, there may be left censoring so we devised a "missingness" model to simulate the possibility that a peptide could be present in a sample but not observed.

To assess the accuracy of the procedure in determining protein presence (Step 1), we evaluated the Type I error rate, the proportion of times that the method correctly assigned p-values ≤ 0.01 to the target fraction, and the proportion of the time that it correctly classified the fraction as either target or non-target. To evaluate the final protein assignment procedure (Step 2), we performed three tests. First, we calculated the probability that the fraction with the highest estimated abundance, the target location, was indeed assigned. Next, we computed the percent of the time that the target fraction's p-value was ≤ 0.01 and the target fraction was included in the final assignment. And finally, to gain perspective on the precision of this procedure, we calculated the average number of times that more than one fraction was included in the final assignment when only one fraction should have been included. The details of the simulation study are given next.

2.2.3.1 Simulation study details

For simplicity, we allowed for only one peptide, j, in any fraction, with a total number of samples equal to 9. The observations from the target fraction are distributed differently from the observations of the "non-target" fractions. For the target fraction, let $\omega_{ik} \sim N(10, 2)$ and $\sigma_{ik} \sim Exp(1)$. Then, for protein i in the target fraction, $Y_{ijks} \sim N(\omega_{ik}, \sigma_{ik})$ and for protein i in any of the other fractions, $Y_{ijks} \sim N(3, 1)$.

Additionally, we must take left censored data into account. For this type of data, the lower the peptide abundance, the more likely it is that the peptide is not observed in the sample. Recall that $Z_{ijks} = 1$ if a peptide is observed; otherwise, $Z_{ijks} = 0$. In our missingness model, $P(Z_{ijks} = 1)$ is a function of Y_{ijks} so that higher values of Y_{ijks} produce higher probabilities of being observed. Then, $Z_{ijks} \sim$ Bernoulli($P(Z_{ijks} = 1)$). The function used to generate these probabilities, illustrated in Figure 2.1, is a scaled logistic($\mu = 5, s = 2$) cumulative distribution function.

$$P(Z_{ijks} = 0 | Y_{ijks} = y_{ijks}) = 1 - \frac{1.08}{1 + e^{-(y_{ijks}-5)/2}} + \frac{1}{1 + e^{-(-5/2)}}$$

We evaluated the accuracy of the method in terms of six tests.

Test 1: To assess the Type I error rate, we calculated the proportion of times we decided the protein was present in a fraction but it was actually not present.

Test 2: In order to assess the accuracy, we calculated the proportion of the time that p-values ≤ 0.01 were correctly assigned to the target fraction.

Test 3: To assess the overall accuracy, we calculated the proportion of the time that we correctly assigned the proteins to the target fraction and correctly assigned proteins not in the target fraction to the non-target fractions.

To test the soundness of the method in making the final assignment (Step 2),

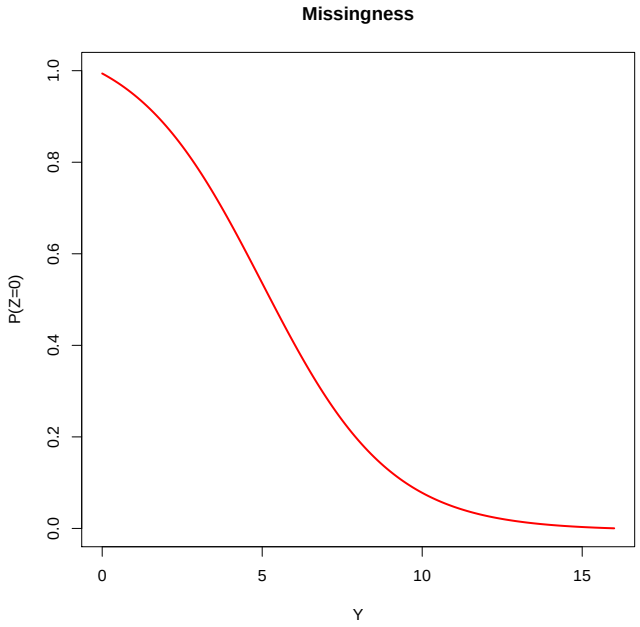

Figure 2.1: Illustration of the "missingness" function used in the simulation study. The x-axis represents the peptide abundance and the y-axis represents $P(Z_{ijks} = 0|Y_{ijks} = y_{ijks})$.

it is important to understand that parts of the results of these tests depend on the results of the previous tests. First, we filtered the data based on the p-values from the protein presence test and only considered those fractions with p-values ≤ 0.01 as candidates for $\hat{G}_i$. If a protein was not present in a fraction, it was considered not to be in $\hat{G}_i$. Recall that

$$\hat{G}_i = \{k : \hat{\omega}_{ik} \text{ are not significantly different}\}$$

and that at the $\alpha_2 = 0.01$ level, the probability that the largest abundance should be in G_i is $1 - \alpha_2 = 0.99$.

Test 4: First, we considered only the target fractions and calculated the proportion of the time that the target fraction was in $\hat{G}_i$. This assessed the probability

that the fraction with the highest simulated abundance (the target fraction) was considered "best". This proportion should be greater or equal to 0.99.

Test 5: To evaluate the overall correctness of both the presence and final assignment we computed the proportion of the time that the target fraction was in $\hat{G}_i$.

Test 6: We also tested the precision of the procedure as follows: In the previous tests, a success was defined as the target fraction being in $\hat{G}_i$ but it was possible that there was more than one fraction in $\hat{G}_i$ and it is important to determine, on average, how many times this occurred. Thus, we calculated the average number of times that there was more than one fraction in $\hat{G}_i$ when there should have only been one fraction in $\hat{G}_i$ and called this measure the precision of the test.

2.3 Results and discussion

2.3.1 Proteins resolved to a single subcellular fraction

Our method is particularly informative where peptides for a protein are measured in multiple fractions, but the likely fraction of localization is not apparent. For example, unique peptides for 483 proteins were identified in multiple subcellular fractions for *R. sphaeroides* 2.4.1 cultured under non-photosynthetic conditions. Of these, there was not significant evidence (at the $\alpha = 0.01$ level) of protein presence in any fraction for 131 proteins but the primary fraction of localization was resolved for 270 proteins (56%). When we relaxed the Type I error rate to the $\alpha = 0.05$ level, 334 of the 483 proteins (69%) for which unique peptides are observed in multiple fractions were resolved to a primary fraction. Similarly, for the photosynthetic condition, there was not significant evidence at the $\alpha = 0.01$ level to conclude that the protein was present in any fraction for 162 of the 559 proteins but the primary fraction for 249 of 559 (45%) proteins were resolved. When we increased $\alpha = 0.05$, 318 of the 559 proteins (57%) in the photosynthetic condition were resolved to a primary fraction.

Results from our statistical assignment of proteins to subcellular fractions were compared to previous non-statistical assignments. Zeng et al., 2007 (Zeng et al., 2007) focused on proteomes of the intra-cytoplasmic membrane (ICM), cytoplasmic membrane (CM), and outer membrane (OM) fractions in order to better characterize proteins associated with the ICM. Ten proteins were classified as ICM unique, while another 18 were classified as ICM associated (larger number of peptides observed in the ICM, but peptides also the CM), and 42 proteins were classified as CM enriched (larger number of peptides observed in the CM, but peptides also observed in the ICM).

Our results agreed with the assignment of well characterized photo apparatus proteins such as PufM (RSP0256), PufL (RSP0257), and pigment proteins (RSP0314, RSP1556) as ICM unique. Welch's t-test for the photo reaction center protein (RSP0291) could not reject the null-hypothesis (no significance difference in its abundance between the ICM and CM), so it was assigned to both subcellular fractions.

However, our analysis could not validate the observation made by Zeng et al., 2007 (Zeng et al., 2007) concerning RSP6124 (Conserved hypothetical), RSP1760 (Conserved hypothetical) and RSP3246 (Putative D-alanyl-D-alanine carboxypeptidase) as ICM unique. While peptides for RSP6124 were confidently identified in all 5 subcellular fractions, Hsu's procedure (at the $\alpha = 0.01$ level) assigned this protein to the cytoplasm. For RSP1760, Hsu's procedure could not resolve the primary subcellular fraction among the OM, ICM, and CM. This is in contrast to Western blot evidence that showed greater banding intensity of histidine tagged RSP1760 in the ICM compared to the CM (Zeng et al., 2007). Finally, our analysis confidently identified peptides for RSP3246 only in the cytoplasm and periplasm, with a statistically greater abundance of this protein occurring in the cytoplasm.

The ability to resolve the primary fraction of localization using this method allows

for the development of biological hypotheses related to specific proteins. Peptides for RSP0842, annotated as a putative porin protein (BLAST result against SwissProt), were identified in multiple subcellular fractions for both *R. sphaeroides* 2.4.1 non-photosynthetic and photosynthetic cell states. In the non-photosynthetic cell state, the estimated standard error associated with the estimated abundance of RSP0842 in each fraction resulted in not assigning the protein to any fraction. However, under conditions that induce photosynthesis, we assigned (at the $\alpha = 0.01$ level) RSP0842 to the intracytoplasmic membrane (ICM; p-value = 0.007) and did not assign it to the periplasm (p-value = 0.185), cytoplasmic membrane (p-value = 0.045), or outer membrane (p-value = 0.022) fractions. Porin proteins are largely considered to be outer membrane proteins that facilitate the passage of small hydrophilic molecules. When we increased α from 0.01 to 0.05, the p-value corresponding to the outer membrane is significant and the method assigns the protein to the outer membrane, cytoplasmic membrane, and ICM. To our knowledge, the association of RSP0842 with the ICM has not been reported for *R. sphaeroides* 2.4.1 and follow-up experimentation to confirm this statistical based observation is needed.

2.3.2 Comparison to PSORTb

We applied our method to the data collected on *R. sphaeroides* 2.4.1 and compared the results to PSORTb's predictions. PSORTb v.3.0 predicts either a single fraction or "Unknown" if the protein may be present in multiple fractions. In this case, it is recommended that the long form of the PSORTb analysis be studied to determine a prediction. For gram negative proteins, the possible final predictions for PSORTb are "Unknown", "Cytoplasmic", "CytoplasmicMembrane", "Periplasm", "OuterMembrane", and "Extracellular". In both conditions, our method may return any combination of "Cytoplasmic", "CytoplasmicMembrane", "Periplasm", and

"OuterMembrane". Proteins in the photosynthetic condition may also have the assignment of "ICM" which stands for intra-cytoplasmic membrane. Our method returned assignments for 567 proteins in at least one condition, either the non-photosynthetic or photosynthetic condition, or both. Of these 567, PSORTb returned "Unknown" for 110 (19%) proteins. We compared our assignments to PSORTb's predictions for the remaining 457 proteins.

There were 384 proteins in the non-photosynthetic condition. Our method and PSORTb agreed exactly for 243 (63%) of the proteins. However, there were an additional 62 proteins for which we assigned multiple fractions, with PSORTb's prediction included in our assignment. Thus, there was at least some overlap for 305 proteins (79.4%).

In the photosynthetic condition, there were 376 proteins for which both PSORTb and our method returned an assignment or prediction. Of these proteins, our method assigned the exact same fraction as PSORTb for 202 proteins (54%). For those proteins that we assigned to have multiple fractions, PSORTb's prediction was included in our assignment for 113 proteins. This means there was at least some overlap in the predictions for 315 proteins (84%) in the photosynthetic condition.

Overall, there were 317 of the 457 proteins (69.4%) that our assignment matched PSORTb's prediction exactly in either the non-photosynthetic or photosynthetic condition and our assignment for 387 proteins (84.7%) had at least some overlap with PSORTb's prediction in either the non-photosynthetic or photosynthetic condition. There were 303 proteins for which we returned an assignment in both conditions and only 33 (11%) of these proteins did not have some overlap with PSORTb in at least one condition.

We randomly chose a few proteins for which we made the same assignment under both conditions but our assignment differed from PSORTb's prediction. Figure 2.2

and Figure 2.3 correspond to the protein ABA79306 under the non-photosynthetic and photosynthetic conditions, respectively. Our method assigned the protein to the periplasm in both conditions, while PSORTb predicted the protein was in the cytoplasmic membrane. In the non-photosynthetic condition (Figure 2.2), there were 50 observations of peptides corresponding to this protein in the periplasm and 27 observations of peptides corresponding to this protein in the cytoplasm but no observations of peptides corresponding to this protein in the cytoplasmic membrane. In the photosynthetic condition (Figure 2.3), there were 44 observations of peptides in the periplasm and 2 observations of peptides in the cytoplasm but no observations of peptides corresponding to this protein in the cytoplasmic membrane.

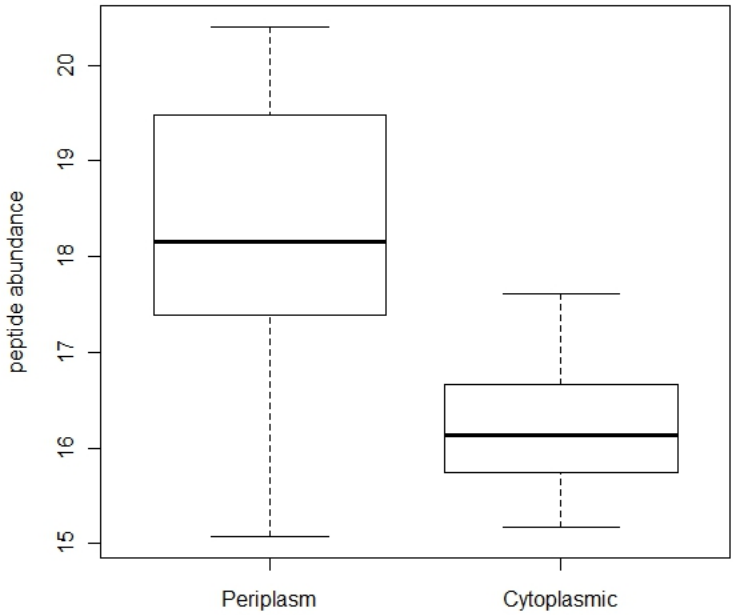

Figure 2.2: Boxplots of the observed peptide abundance for various fractions corresponding to protein ABA79306 under the non-photosynthetic condition. There were 50 observations in the periplasm and 27 observations in the cytoplasm.

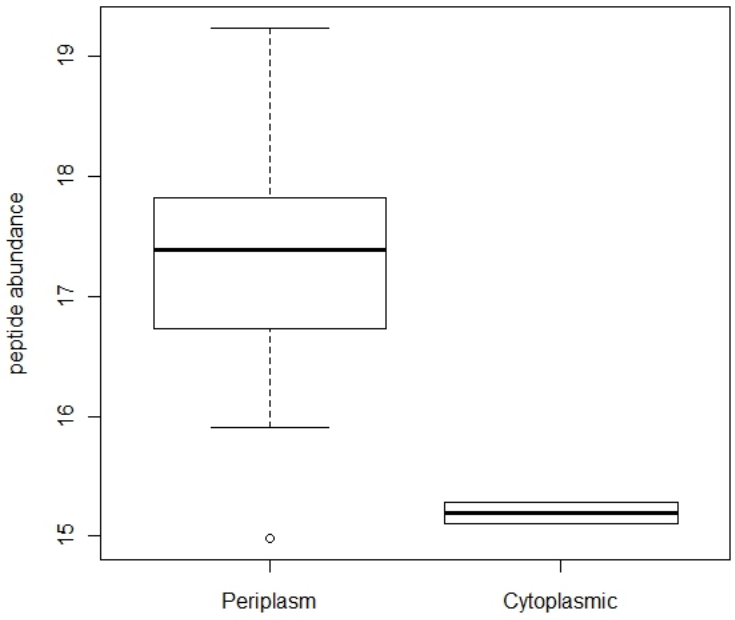

Figure 2.3: Boxplots of the observed peptide abundance for various fractions corresponding to protein ABA79306 under the photosynthetic condition. There were 44 observations in the periplasm and 2 observations in the cytoplasm.

Figure 2.4 and Figure 2.5 correspond to the ABA78029 protein. We assigned the protein to the outer membrane under both conditions, while PSORTb predicted the protein is located in the cytoplasmic membrane. Under the non-photosynthetic condition (Figure 2.4), there were 3 observations of peptides in the periplasm, 5 observations of peptides in the cytoplasmic membrane, and 15 observations of peptides in the outer membrane. Under photosynthetic conditions (Figure 2.5), there were 4 observations of peptides in the cytoplasmic membrane and 20 observations of peptides in the outer membrane. Furthermore, the boxplots of the peptide abundance illustrate that the peptide abundances observed in the cytoplasmic membrane are much smaller than the peptide abundances observed in the outer membrane.

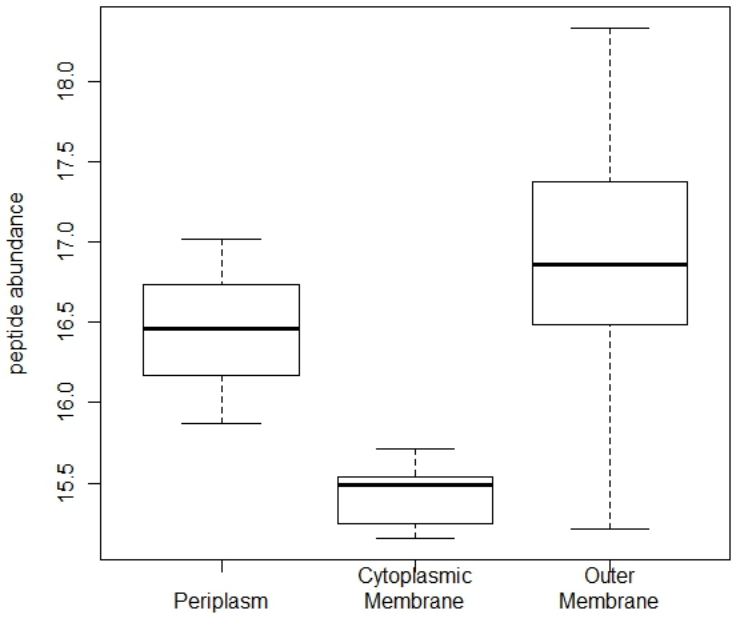

Figure 2.4: Boxplots of the observed peptide abundance for various fractions corresponding to protein ABA78029 under the non-photosynthetic condition. There were 3 observations in the periplasm, 5 observations in the cytoplasmic membrane, and 15 observations in the outer membrane.

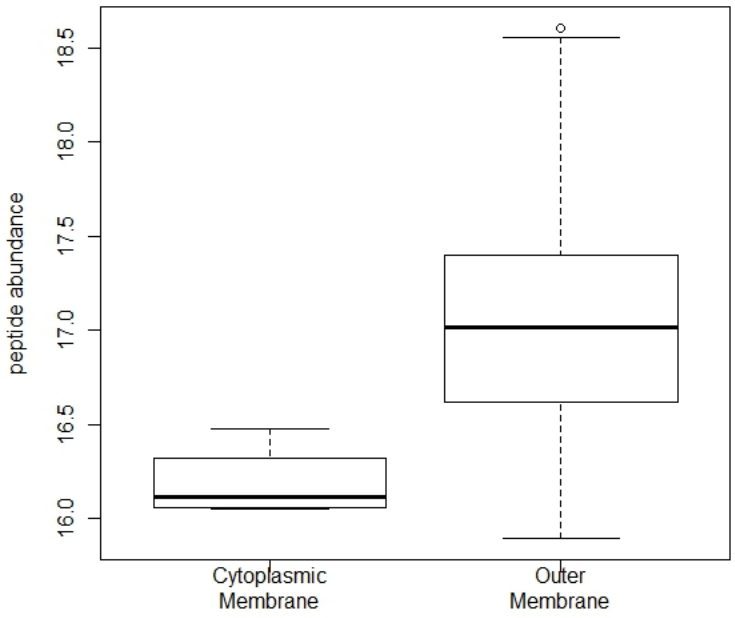

Figure 2.5: Boxplots of the observed peptide abundance for various fractions corresponding to protein ABA78029 under the photosynthetic condition. There were 4 observations in the cytoplasmic membrane and 20 observations in the outer membrane.

2.3.3 Simulation study results

The simulation study revealed that the method produces sound and accurate results for the simulated data. We used Wilson Score confidence intervals on all of the proportions because its coverage better approaches the nominal value for small samples than the usual Wald confidence interval. The Wilson Score interval is computed as follows:

$$\frac{\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{z^2_{1-\alpha/2}}{n}},$$

where $\hat{p}$ is the sample proportion, n is the number of samples, and $z_{1-\alpha/2}$ corresponds to $1 - \alpha/2$ percentile of the standard normal distribution.

When assessing the procedure that determines protein presence in a fraction, the Type I error rate at the $\alpha = 0.01$ level was 0.006 with a 99% Wilson Score confidence interval of [0.004,0.015]. The proportion of the p-values ≤ 0.01 in the target fraction was 0.926 with a 99% Wilson Score confidence interval of [0.895, 0.957]. Additionally, we found that the proportion of p-values ≤ 0.01 in the target fraction and the p-values > 0.01 in the non-target fractions was 0.980 with a 99% Wilson Score confidence interval of [0.973, 0.988].

Likewise, tests on the final assignment procedure produced excellent results. For the first test, the data were filtered based on the presence of a protein in a fraction at the $\alpha = 0.01$ level. That is, a protein cannot be considered to have the largest abundance if it is not present in the fraction at a significant level. When the data were first filtered, the proportion of times that the target fraction was included in the final assignment was 1.00 with a 99% Wilson Score confidence interval of [1, 1]. In estimating the overall success rate, we found that the proportion of times the method correctly identified both presence of a protein and the protein's correct fraction was 0.926 with a 99% Wilson Score confidence interval of [0.897, 0.957]. Lastly, the proportion of times that the final assignment contained more than one fraction when it should only contain one fraction was 0.00 with a 99% Wilson Score confidence interval of [0.000, 0.026]. This last result indicates that the procedure accurately classifies the target fraction as being in the final assignment set but it does not spuriously include non-target fractions in the final assignment set.

2.4 Conclusions

Our method is very flexible. It can be applied to any organism for which this type of data are available. It also compared very well with PSORTb. Overall, our assignments matched PSORTb's prediction exactly in either the non-photosynthetic or photosynthetic condition 69% of the time and at least partially matched PSORTb's prediction in at least one condition for 85% of the proteins. Additionally, our method was able to assign a fraction for approximately 19% more proteins than PSORTb.

The results of the simulation study further confirmed that the methods outlined in this paper are sound and produce very good results. The Type I error rate's confidence interval includes 0.01 and exceeds 0.01 by 0.005, an acceptable margin. The proportion of times that correct p-values were assigned was 0.98 with a lower bound of 0.973. Furthermore, combining the protein presence estimations with Hsu's procedure gave a proportion of correctly identifying the presence of a protein and its fraction of 0.926, with a lower bound of 0.897. This implies that 90% of the time, the method got it right. Another important facet of Hsu's procedure is that more than one fraction can be included in the final assignment. For this reason, it is important to test the precision of our final assignment and make sure that it did not spuriously include other fractions. This event never occurred in the simulation study and a 99% confidence interval had the upper limit of the probability of this event occurring at 0.026 (2.6%). Thus, not only is this method accurate, it is also precise.

3. PREDICTING THE LETHAL PHENOTYPE OF THE KNOCKOUT MOUSE BY INTEGRATING COMPREHENSIVE GENOMIC DATA*

3.1 Introduction

The mouse is the premier model organism for interpreting the human genome and plays a key role in studying human diseases (Collins et al., 2007). Importantly, the mouse is the only vertebrate species in which pre-selected genes can be deliberately mutated (knocked out) such that the phenotypic effect associated with a gene can be defined in a precise manner. Among various phenotypic effects of disrupting a mouse gene, the lethal phenotype is of particular interest.

The primary biological questions we aimed to address were as follows:

1. Which genomic features are most important to predicting the lethal phenotype of mouse single-gene knockouts?

2. Through reasonable computational approaches, to what extent can the knockout lethality be predicted from a wide range of genomic features?

This study compared the accuracy of three classification methods: logistic regression, support vector machine (SVM), and random forest. SVM and random forest are machine learning methods that can be used to classify a binary response. In order to fairly compare the methods, we used least absolute shrinkage and selection

*Part of this section is reprinted with permission from "Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data" by Yuan, Y., Xu, Y., Xu, J., Ball, R.L., and Liang, H., 2012. *Bioinformatics*, 28, 1246–1252. Copyright 2012 by Yuan, Y., Xu, Y., Xu, J., Ball, R.L., and Liang, H.

operator (LASSO) penalty (Tibshirani, 1996) for feature selection and allowed all methods the same set of features. All the features with non-zero coefficients were included in feature set.

My role in this endeavor was to apply the random forest technique to the data using 5-fold cross validation, using LASSO for feature selection, and performing a study bias correction when assessing the accuracy of the classifiers.

3.2 Methods

3.2.1 Data

Phenotype information was collected from the Mouse Genome Informatics (MGI) database and the gene coding sequence, protein domain, gene homology and structural information were downloaded from Ensembl (release 59). There were 4670 genes with a known response (lethal or nonlethal) and 15,175 genes without a known response. We designated the 4670 genes with a known response as the training set and the remaining 15,175 genes as the predicting set. For each gene, there were 491 genomic features (Yuan et al., 2012).

3.2.2 Statistical methods

A random forest is a collection of decision trees such that each tree is built from a random subset of the data. The random forest technique was first introduced in 2001 by Leo Breiman (Breiman, 2001), and since then it has been shown to be a highly accurate classifier in a number of fields, including genetics (Bureau et al., 2005). We used the 'randomForest' package in R (Liaw and Wiener, 2002) and chose parameter values according to Breiman's methodology. Although the default number of trees is 500, we chose to build 5000 trees (ntree = 5000) to obtain more robust results. Each tree was grown to its full depth (nodesize = 1) and was not pruned. At each node of each tree, a different random subset of the features was selected, and the Gini

criterion was used to determine the feature in this subset that produced the best split of the data. The size of this subset (mtry) was the square root of the number of possible features. The other parameters were set as defaults.

We used LASSO for feature selection by utilizing the R package 'glmnet' (Friedman et al., 2010). The value of the tuning parameter, λ, was obtained through a cross-validation procedure. If the feature had a non-zero coefficient, it was included in the feature set to be used in each of the classification methods. An overall scheme of the project is given in Figure 3.1.

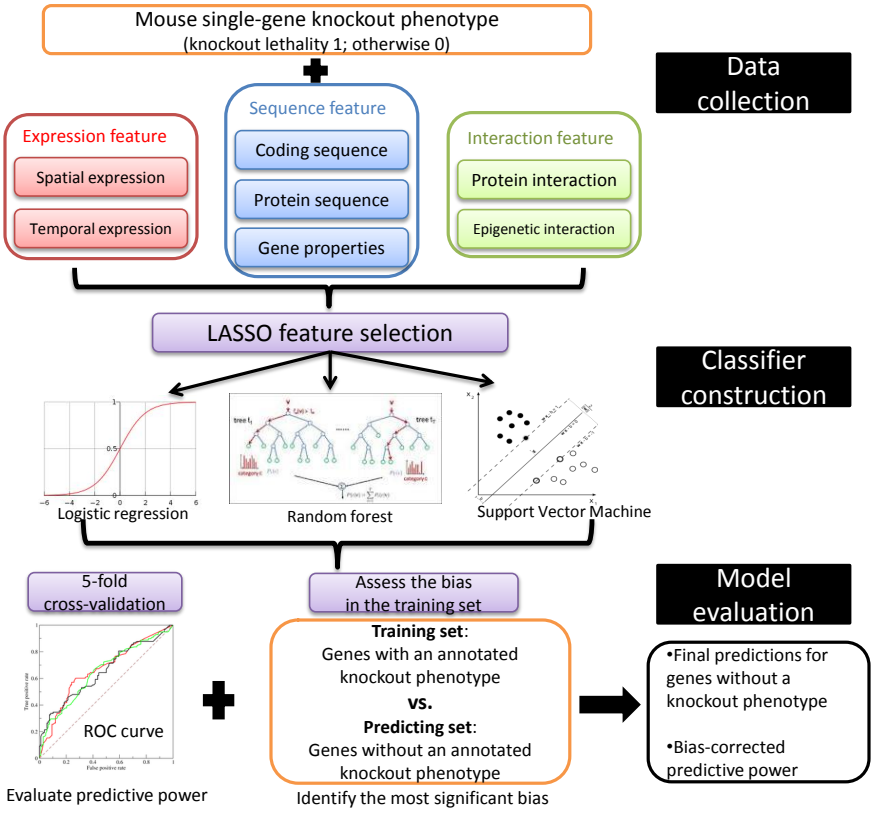

Figure 3.1: Overall scheme of the project.

The random forest classifier performed best in terms of area under the ROC curve, so it was used to make the final predictions for the genes. Since the random

forest predictions were probability predictions, we chose the cutoff probability for determining the final classification (1 for lethal, 0 for nonlethal) by maximizing the accuracy.

We also needed to address the possible study bias in the training set. Some genes may be studied more often than others because of the results of previous studies or known interactions. Thus, the training set may contain more frequently studied genes and we need to take this into account. To assess the possible study bias in the training set, we computed the mutual information of features across the training and predicting sets. We found that the training set was particularly biased in terms of the evolutionary age feature. Ancient genes were more likely to be studied. Thus, when computing the various accuracy measures, the measures were weighted by the proportion of the evolutionary age group in the predicting set. There are seven groups with the first group consisting of genes with evolutionary ages 1-6 and the last group comprised of genes with an evolutionary age equal to 12.

Let p_i be the proportion of evolutionary age group i in the predicting set. From the 2×2 contingency table, we have the number of positives (P), the total number of negatives (N), the number of true positives (TP), the number of false negatives (FN), the number of false positives (FP), and the number of true negatives (TN). Then, accuracy (Acc), positive predictive value (PPV), negative predictive value (NPV),

and recall are defined as follows:

$$\begin{aligned} \text{Acc} &= \frac{\text{TP+TN}}{\text{P+N}} \\ \text{PPV} &= \frac{\text{TP}}{\text{TP+FP}} \\ \text{NPV} &= \frac{\text{TN}}{\text{TN+FN}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP+FN}} \end{aligned}$$

We computed each measure for each evolutionary age group i and took the weighted measures of accuracy as follows:

$$\begin{aligned} \text{Acc*} &= \sum_i p_i \text{Acc}_i \\ \text{PPV*} &= \sum_i p_i \text{PPV}_i \\ \text{NPV*} &= \sum_i p_i \text{NPV}_i \\ \text{Recall*} &= \sum_i p_i \text{Recall}_i \end{aligned}$$

3.3 Results

The gene features were categorized into three different groups:

- The genomic sequence set (S) is comprised of 373 features pertaining to coding sequences, protein sequences, and other properties such as gene regulation.
- The mRNA expression set (E) is comprised of 94 features pertaining to tissue expression profiles and developmental stage profiles.

- The interaction set (I) is comprised of 24 features that pertain to mouse protein-protein interaction and mouse epigenetic histone modification interactions.

Table 3.1 lists the top 20 features selected by LASSO based on the S+E+I gene feature set. The feature with the largest coefficient is evolutionary age. In the mRNA expression set (E), expression level in utero and expression level in development stage 15 ranked highest. Protein connectivity had the largest coefficient in the interaction set (I). Although some of the features selected also exhibited a high correlation with lethality, others, such as paralog sequence identity, did not show a high correlation with knockout lethality.

Feature	LASSO coefficient
Evolutionary age	0.473
Expression in utero	0.436
Expression in TS15	0.297
Paralog sequence identity	-0.296
Total miRNA target sites	0.254
Expression in TS11	0.196
Connectivity	0.160
Expression in TS17	0.146
Expression in TS26	0.132
Expression in TS18	0.129
Total histone modification	0.128
Asparagine content	0.113
5-UTR length	-0.110
Expression in upper spinal cord	-0.110
Leucine content	-0.106
Expression in bone	-0.104
Expression in TS5	0.100
Amino acid length	0.093
Expression in ovary	0.089
Expression in TS19	0.086

Table 3.1: Top 20 informative genomic features related to knockout lethality, as selected by LASSO

Using only interaction features (I), we achieved a low AUC=0.598 but sequence features alone (S) returned an AUC=0.738 and expression features alone (E) gave an AUC=0.725. As more features were added to the classifiers, the AUC increased with the highest AUC=0.782 achieved when all features were candidates for inclusion (S+E+I). Without correcting for the study bias of the evolutionary age feature in the training set, the accuracy for the S+E+I random forest classifier was 72.5% and recall was 62.9%. When we corrected for the study bias (as described in the Methods section), we achieved an accuracy (Acc*) of 70.9% and recall equal to 60.7%. Tables 3.2 and 3.3 summarize the performance of the random forest classifier across different feature sets. Table 3.2 gives the values before correcting for the study bias in the evolutionary age feature in the training set and Table 3.3 gives the values after we correct for the study bias in the training set.

Feature set	# Features included	Cut-off	Acc	PPV	NPV	Recall
S	36	0.48	0.709	0.624	0.737	0.438
E	24	0.49	0.705	0.615	0.734	0.432
I	20	0.58	0.644	0.598	0.650	0.184
E+I	27	0.57	0.697	0.720	0.693	0.323
S+E	57	0.54	0.727	0.703	0.733	0.391
S+I	37	0.49	0.706	0.636	0.738	0.528
S+E+I	44	0.46	0.725	0.637	0.777	0.629

Table 3.2: Results for the various feature sets without correcting for study bias. Cut-off was the value at which the probability prediction classified the gene as lethal. Everything above cut-off was considered lethal. Accuracy (Acc), positive predictive value (PPV), negative predictive value (NPV) and Recall were defined previously.

3.4 Discussion and conclusion

We compared the performance of three classifiers: logistic regression, support vector machine, and random forest. We found that the random forest performed best in terms of AUC and we achieved the best results (Acc=0.725) on the S+E+I feature set, which included 44 of 491 features in the final classifier. Because mouse

Feature set	# Features included	Cut-off	Acc*	PPV*	NPV*	Recall*
S	36	0.47	0.744	0.568	0.756	0.331
E	24	0.55	0.720	0.599	0.725	0.296
I	20	0.49	0.628	0.551	0.634	0.291
E+I	27	0.54	0.679	0.670	0.667	0.371
S+E	57	0.54	0.745	0.646	0.740	0.315
S+I	37	0.49	0.696	0.600	0.706	0.499
S+E+I	44	0.46	0.709	0.642	0.738	0.607

Table 3.3: Results for the various feature sets when we corrected for the study bias in the training set. Cut-off was the value at which the probability prediction classified the gene as lethal. Everything above cut-off was considered lethal. Accuracy (Acc*), positive predictive value (PPV*), negative predictive value (NPV*) and Recall* were defined previously.

knockout experiments are so time-consuming, we hope that scientists will use our method to choose the best genes to investigate further.

We also investigated a large number of features that had not previously been studied. We found that evolutionary age was consistently chosen as one of the top features for predicting lethality. This suggests that disrupting an evolutionary ancient gene results in more severe consequences. Also, although paralog sequence identity did not have a high correlation with knockout lethality, it was also included in the model because of a large negative LASSO coefficient. This suggests that if a gene is removed and that gene has a closely related copy that is not removed, there is not likely to be a lethal consequence. Other interesting features related to gene expression, such as expression level in utero and TS15, were also identified and deserve further study (Yuan et al., 2012). The paper (Yuan et al., 2012) discusses these findings in more depth.

4. COMPARATIVE ANALYSIS OF SOMATIC COPY-NUMBER ALTERATIONS ACROSS DIFFERENT HUMAN CANCER TYPES REVEALS TWO DISTINCT CLASSES OF BREAKPOINT HOTSPOTS*

4.1 Introduction

Copy-number variations occur when parts of the cell's DNA develop an abnormal number of copies, either too many or too few. Somatic copy-number alterations (SCNAs) play a significant role in the development of human cancers and studying SCNAs can lead to important discoveries of cancer-causing genes and the development of possible treatment strategies.

DNA breakpoints are defined as breaks in the chromosome that later recombine. Breakpoints that are clustered in particular regions of the human genome are called breakpoint hotspots, or hotspots. In this study, we consider data from eight different cancer types (listed in Table 4.1) to investigate cancer-type specific hotspots and common hotspots. Cancer-type specific hotspots are identified in one or two of the eight cancer types and common hotspots are identified in at least seven of the eight cancer types studied. Previous studies focused on pooled data sets from different cancer types or only on a single cancer type. We utilized data from The Cancer Genome Atlas (TCGA) project to perform a comparative analysis of breakpoints across eight cancer types. The cancer types are listed in Table 4.1 and summary

*Part of this section is reprinted with permission from "Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots" by Li, Y., Zhang, L., Ball, R.L., Liang, X., Li, J., Lin, Z., and Liang, H., 2012. *Human Molecular Genetics*, 21, 4957–4965. Copyright 2012 by Li, Y., Zhang, L., Ball, R.L., Liang, X., Li, J., Lin, Z., and Liang, H.

statistics on the number of breakpoints for each cancer type are given in Table 4.2.

Cancer type	Abbreviation	Number of samples
Breast invasive carcinoma	BRCA	667
Glioblastoma multiforme	GBM	441
Ovarian serous cyst adenocarcinoma	OV	500
Kidney renal clear cell carcinoma	KIRC	459
Colon adenocarcinoma	COAD	892
Uterine corpus endometrioid carcinoma	UCEC	272
Lung squamous cell carcinoma	LUSC	187
Lung adenocarcinoma	LUAD	163

Table 4.1: Eight cancer types investigated in this study.

Cancer type	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
BRCA	0	11	17	40.22	27	1546
GBM	0	4	7	21.46	13	836
OV	1	16	23	41.33	34.75	1021
COAD	0	2	4	17.97	8	819
KIRC	0	0	2	16.15	4	970
UCEC	0	2	4	13.04	8	524
LUSC	0	2	4	10.47	7	426
LUAD	0	1	2	7.58	5	338

Table 4.2: Summary statistics on the number of breakpoints for eight cancer types investigated in this study.

My role in this study was to use the random forest technique to investigate the predictability of cancer-type specific hotspots and common hotspots using genomic features and to investigate the relationship between breakpoint frequency and the genomic features using multivariate regression analysis.

4.2 Methods

The random forest classifiers were implemented using the 'randomForest' package (Liaw and Wiener, 2002) in R according to Breiman's methodology (Breiman, 2001). Although the default number of trees is 500, we chose to build 1000 trees (ntree=1000) to obtain more robust results. Each tree was grown to its full depth

(nodesize=1) and was not pruned. At each node of each tree, a different random subset of the features was selected, and the Gini criterion was used to determine the feature in this subset that produced the best split of the data. The size of this subset (mtry) was the square root of the number of possible features (mtry=$\sqrt{p}$, where p is the number of features). Otherwise, the parameter values were left at their default values. To assess the predictive power, we performed 10-fold cross-validation: in each round, 90% of the data were used as the training data, and the remaining 10% were used as the test data.

To investigate the relationship between the various genomic features and breakpoint frequency, we first transformed the breakpoint frequency so that it was approximately normally distributed. If y_i is the breakpoint frequency, the transformed breakpoint frequency is defined as ${y_i}^t = \frac{y_i}{y_i+\tilde{\mu}}$ where $\tilde{\mu}$ is the median of the sample. We randomly divided the data into a training and test set with 70% of the data assigned to the training set and 30% of the data assigned to the test set. We used forward stepwise regression to investigate the relationship between the 19 genomic features and breakpoint frequency. The 19 genomic features included sequence features, DNA structural motifs, evolutionary features, and functional features.

4.3 Results

Table 4.3 summarizes the results of the forward selection regression analysis. We computed the adjusted R^2 over the training data and the R^2 over the test data at each step in the forward regression. The features were added in the order presented. "All" is the sum of breakpoints across cancer types. The features that were selected the most were indel rate, substitution rate, exon density, and sine. These features are also shown to have high variable importance measures in the random forest (Figure 4.2 and Figure 4.3).

Cancer type	Features Included	Adj R^2 (Training)	R^2 (Test)
All	indelRate + exon + subRate + sine + line + recRate + gc	0.223	0.156
BRCA	subRate + sine + exon + recRate + gc + line	0.143	0.111
GBM	subRate + exon + sine + fra + line + gc	0.154	0.103
OV	indelRate + exon + subRate + sine + line + slipped	0.235	0.153
KIRC	subRate + triplex + recRate + exon	0.073	0.103
COAD	subRate + exon + gc + cons17way	0.099	0.080
UCEC	exon + subRate + sine + recRate + cruciform + line + repTime + zdna	0.218	0.171
LUSC	subRate + indelRate + gc + recRate + sine + cruciform	0.099	0.066
LUAD	indelRate + subRate + exon	0.073	0.064

Table 4.3: Results from the forward selection regression on the transformed breakpoint data.

There were a total of 2822 hotspots. Of these, 138 were classified as common hotspots (flag = 7 or 8) and 217 were classified as cancer-type specific hotspots (flag = 1 or 2). We built two random forest classifiers. The first attempted to distinguish between a common hotspot and all other hotspots. The second distinguished between a cancer-type specific hotspot and all other hotspots. Figure 4.1 displays the ROC curve of the resulting classifiers. The area under the curve (AUC) is 0.615 for the cancer-type specific hotspot classifier and 0.748 for the common hotspot classifier. The random forest classified common hotspots better than it classified cancer-type specific hotspots. The resulting variable importance plots are shown in Figures 4.2 and 4.3. These plots give a ranking of each feature's relative importance in predicting the response.

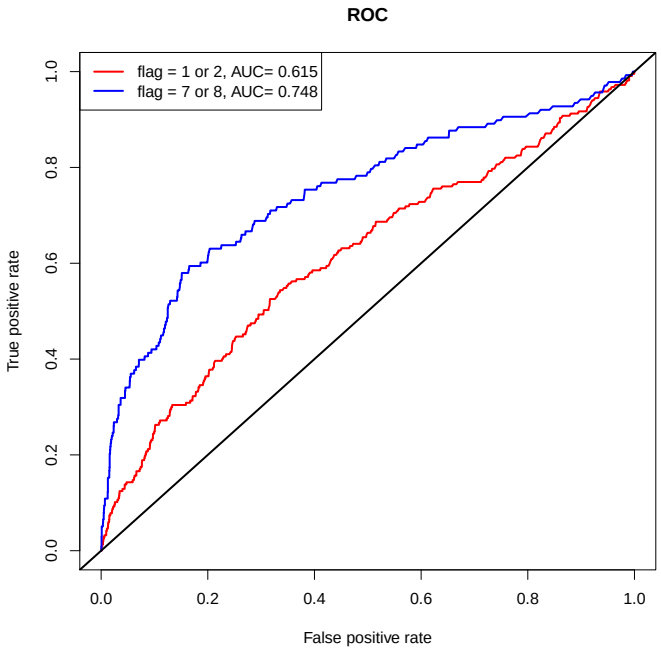

Figure 4.1: ROC for the two random forest classifiers. Common hotspots have flag=7 or 8 and cancer-type specific hotspots have flag=1 or 2

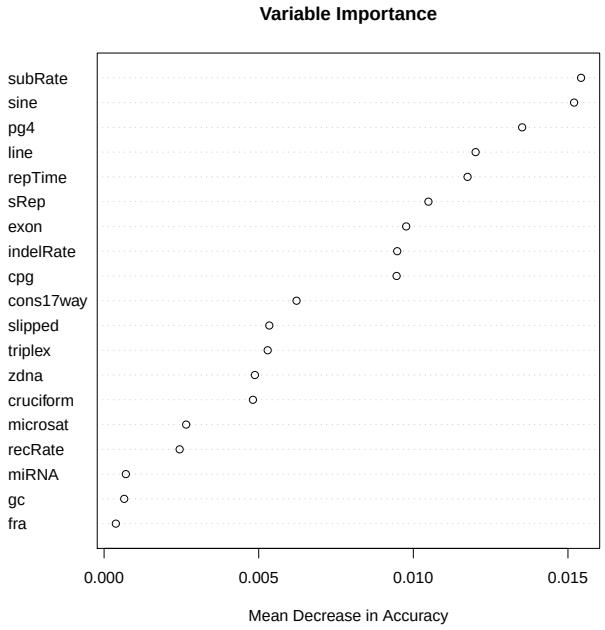

Figure 4.2: Variable importance plot for cancer-type specific hotspots. Larger mean decrease in accuracy values indicate more important features.

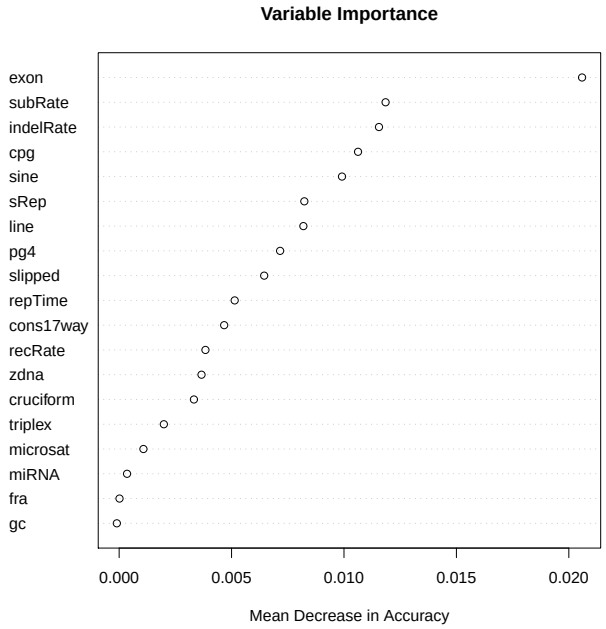

Figure 4.3: Variable importance plot for common hotspots. Larger mean decrease in accuracy values indicate more important features.

4.4 Discussion and conclusion

At the time this study was published, it was the most comprehensive analysis of SCNA data relating to human cancers. It was found that cancer-type specific hotspots are poorly predicted by genomic features, and these hotspots show a significant enrichment for cancer genes, while common hotspots do not have this enrichment property and they can be predicted by the relevant genomic features (Li et al., 2012) . These results highlight the different evolutionary mechanisms present in the common and cancer-type specific hotspots, and these insights are expected to help scientists better understand the critical events in tumorgenesis and progression (Li et al., 2012).

5. PREDICTING "HEART AGE" USING THE ELECTROCARDIOGRAM

5.1 Introduction

Conventional resting electrocardiography (ECG) is currently used as a tool to help diagnose both the form and extent of heart disease. However, its limitations have been well-documented (Ashley et al., 2001; Levy et al., 1990; Sox, Jr et al., 1989). More advanced techniques, such as spatial/spatiotemporal ECG (Kardys et al., 2003) and high-frequency QRS ECG (Schlegel et al., 2004), have been developed that improve the diagnostic power of the ECG (Poplack Potter et al., 2010; Schlegel et al., 2010). The strategy of diagnosis used by Schlegel et al. (2010) uses a linear combination of outputs estimated by logistic regression or discriminant analysis models.

The goal of this study is to predict a subject's hypothetical age based on the ECG-related outputs and a diagnosis of "healthy". We call this hypothetical age "heart age". The question is this: given an individual is healthy (that is, no risk factors) and given the individual's ECG outputs, what is his/her heart age? The problem is that we do not observe heart age; we only observe the ECG outputs and chronological age (body age). Thus, we cannot have a simple regression-type model. We must find some means of inferring heart age based on the observables and prior information. A Bayesian approach is a natural solution.

5.2 Data

Since early 2010, investigators in the Cardiovascular and Neurosciences laboratories at Johnson Space Center have had access to a database (A-ECG) of advanced 12-lead ECG recordings from healthy individuals, individuals with risk factors, and individuals with heart disease. These subjects consisted of volunteers from cardiac

clinics, volunteers from Johnson Space Center and the Universidad de los Andes and Lund University Hospital, subjects who participated in earlier studies at the Charleston Area Medical Center, and healthy subjects from Slovenia (Schlegel et al., 2010). Because of the manner in which subjects were included in this database, the A-ECG database is not a simple random sample from the population.

Each subject was classified according to their cardiac disease status and, if disease was present, then also to the general severity of disease. Disease status was diagnosed based on results from clinical imaging tests (currently, the "gold standard") so that if a subject had heart disease, the form and severity of heart disease was generally known (Schlegel et al., 2010). Subjects were classified as "healthy" if they had no cardiovascular or other systemic disease and also did not have other risk factors, such as hypertension, smoking, or diabetes (Schlegel et al., 2010).

We utilized data from 1,439 subjects that were at least twenty years old, had been given a 5-minute ECG, and were categorized as healthy, diseased, or as having risk factors. Data from another 510 subjects were set aside to be used in subsequent studies. Non-patient volunteers with risk factors, such as diabetes or high blood pressure, were not given a diagnosis.

Subjects were put into 4 groups: healthy non-athletes (HNA), healthy athletes (A), subjects with risk factors and no diagnosis (RFS), and subjects with disease (D). Subjects in the HNA group were ordered based on age and every fourth subject was selected to be in the test set so that 545 subjects were put in the training set and 183 subjects were put into the test set. The athletes were asymptomatic volunteers with no evidence of cardiac disease based on a negative history and physical examination. All were endurance-trained athletes and the majority comprised Swedish triathletes as well as semiprofessional soccer and handball players of both sexes who had cardiac magnetic resonance imaging scans demonstrating no evidence

of hypertrophic cardiomyopathy nor any other clinical pathology (Poplack Potter et al., 2010). Descriptive statistics for these groups are given in Table 5.1.

Group	#subjects	%females	%males	%20–40	%41-60	%60 and older
HNA(train)	545	41%	59%	56%	36%	8%
HNA(test)	183	43%	57%	55%	36%	9%
A	48	38%	62%	92%	6%	2%
RFS	221	47%	53%	10%	59%	31%
D	441	34%	66%	7%	50%	43%

Table 5.1: Descriptive statistics of the groups used in this study.

5.3 Methods

In the Bayesian paradigm, we assume there is a distribution on the parameter of interest and we usually take the mean or the mode of the distribution to be the point estimate of the parameter.

For a given subject, let

$$\begin{aligned} x &= \text{body age} \\ \boldsymbol{y} &= \text{vector of ECG outputs} \\ a &= \text{heart age} \end{aligned}$$

By Bayes Rule, the posterior distribution for heart age is:

$$p(a|x,\boldsymbol{y}) = \frac{p(a|x)p(\boldsymbol{y}|x,a)}{\int p(a|x)p(\boldsymbol{y}|x,a)da}$$

and the predicted heart age, $\hat{a} = E(a|x,\boldsymbol{y}) = \int ap(a|x,\boldsymbol{y})da$, is the mean of the posterior distribution.

Based on expert opinion, we assume that a subject's heart age is approximately normally distributed within 15 years of the subject's body age so the prior distri-

bution of heart age, $p(a|x) \sim N(x, 7.5^2)$, is a normal distribution with mean x and standard deviation $\sigma_a = 7.5$.

Specifying the distribution of $\boldsymbol{y}$ given a and x is a bit trickier. Before including heart age (a) in the distribution, let's look at how we would get $p(\boldsymbol{y}|x)$ when $\boldsymbol{y}$ is one dimensional ($\boldsymbol{y} = y$ and $k = 1$). Assume for the moment that the distribution of y only depends upon x. Based on plots of body age versus ECG outputs, we found that there were often slight nonlinear trends so we assumed a quadratic regression model for the ith healthy subject:

$$y_i|x_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \nu_i; \nu_i \sim N(0, \sigma_\nu^2)$$

for $i = 1, 2, \ldots, n$. So, letting $\vec{y} = (y_1, y_2, \ldots, y_n)^T, \vec{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_\nu^2 \boldsymbol{I})$ where

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}.$$

At first, we considered using weighted least squares to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ because the distribution of ages in our sample is not the same as the distribution of ages in the healthy population. However, provided that the quadratic model holds, the Gauss-Markov Theorem ensures that we obtain an unbiased estimate of $\boldsymbol{\beta}$ with ordinary least squares. Therefore, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$ is the best linear unbiased estimator (BLUE). It follows that an unbiased estimate of σ_ν^2 is $\hat{\sigma}_\nu^2 = \frac{1}{n-3}\sum_{i=1}^n e_i^2$, where $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^T = \vec{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ (Monahan, 2008).

Given the estimates of $\boldsymbol{\beta}$ and σ_ν^2, we incorporate heart age into the model. The

influence of heart age (a) on the distribution of the ECG variable (y) is best illustrated with an example. Consider the case where y increases as x and a increase and suppose we look closely at subjects with a particular body age (keep x fixed). If a were not in the model, we would expect to see a normal distribution of y around $\mathrm{E}(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$. In Figure 5.1, this distribution is represented by a black curve. In the case where heart age is greater than body age, we would expect the mean of y to be shifted right (blue) and if heart age is less than body age, we would expect the mean of y to be shifted left (red). This leads to the following model:

$$(y|x, a) = \beta_0 + \beta_1 x + \beta_2 x^2 + \theta(a - x) + \epsilon,$$

where $\epsilon \sim N(0, \lambda^2)$.

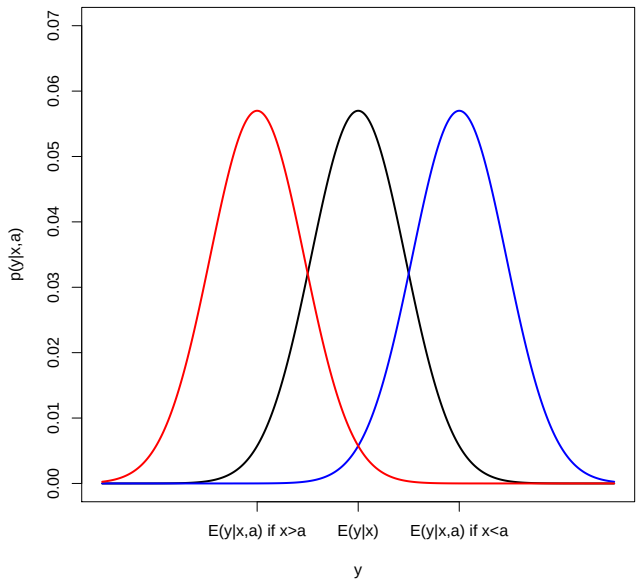

Figure 5.1: Illustration of the effect of heart age on the mean of y.

However, since we do not observe heart age, a, how do we estimate θ? From

estimating $\boldsymbol{\beta}$ using least squares, we obtained an estimate of σ_ν^2 which is, in fact, equal to $\text{Var}\{\theta(a-x)+\epsilon\} = \theta^2\sigma_a^2 + \text{Var}(\epsilon)$ and we have specified that $\sigma_a^2 = 7.5^2$. If we could also estimate $\text{Var}(\epsilon) = \lambda^2$, we could solve for $\theta = \sqrt{\frac{\sigma_\nu^2-\lambda^2}{\sigma_a^2}}$, but we must take the sign of θ into account. Notice that if y and x have a positive increasing relationship, θ should be positive but, if y decreases as x increases, θ should be negative. This is evident by looking at the sign of β_1, therefore,

$$\theta = \text{sgn}(\beta_1)\sqrt{\frac{\sigma_\nu^2 - \lambda^2}{\sigma_a^2}}.$$

A previous study (Batdorf et al., 2006) on the reproducibility and reliability of certain ECG-related outputs was completed by the National Space Biomedical Research Institute and the Human Adaptation and Countermeasures Office at Johnson Space Center in 2005. As a result, we had access to two repeated ECG measures of y, taken a month apart, on $m = 15$ asymptomatic subjects (8 males and 7 females). We assumed that every subject had a fixed heart age which was not likely to change in one month, thus, any variability we observed in y for the subject was due to $\text{Var}(y|x,a) = \lambda^2$. Thus, we used the repeated measures to estimate $\text{Var}(y|x,a) = \lambda^2$ from the differences between the two measurements, noting that $\text{E}(d_i^2) = 2\lambda^2$.

$$\hat{\lambda}^2 = \frac{1}{2m}\sum_{i=1}^{m} d_i^2$$

where $d_i = y_{i1} - y_{i2}$ is the difference between the measurements on the subject i. Recall that we assumed $\sigma_a^2 = 7.5^2$. Therefore, we estimated θ by

$$\hat{\theta} = \text{sgn}(\hat{\beta}_1)\sqrt{\frac{\hat{\sigma}_\nu^2 - \hat{\lambda}^2}{7.5^2}}.$$

5.3.0.1 *Posterior distribution for heart age*

Univariate case Let $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 - \theta x$. The posterior distribution for heart age for a given subject is:

$$\begin{aligned}
p(a|x,y) &= \frac{p(a|x)p(y|x,a)}{\int p(a|x)p(y|x,a)da} \\
&\propto \exp\left\{-\frac{1}{2\sigma_a^2}(a-x)^2\right\} \\
&\quad \times \exp\left[-\frac{1}{2\lambda^2}\left\{y-\beta_0-\beta_1 x-\beta_2 x^2-\theta(a-x)\right\}^2\right] \\
&= \exp\left\{-\frac{1}{2}Q(a)\right\}, \text{ where} \\
Q(a) &= \frac{a^2}{\sigma_a^2} - 2a\frac{x}{\sigma_a^2} + \frac{x^2}{\sigma_a^2} \\
&\quad + \frac{1}{\lambda^2}\left[a\theta - \left\{y - \left(\beta_0+\beta_1 x+\beta_2 x^2-\theta x\right)\right\}\right]^2 \\
&= \frac{a^2}{\sigma_a^2} - 2a\frac{x}{\sigma_a^2} + \frac{x^2}{\sigma_a^2} + \frac{a^2\theta^2}{\lambda^2} - 2a\frac{\theta\{y-g(x)\}}{\lambda^2} + \frac{\{y-g(x)\}^2}{\lambda^2} \\
&= a^2\left[\frac{1}{\sigma_a^2}+\frac{\theta^2}{\lambda^2}\right] - 2a\left[\frac{x}{\sigma_a^2}+\frac{\theta\{y-g(x)\}}{\lambda^2}\right] \\
&\quad + \left[\frac{x^2}{\sigma_a^2}+\frac{\{y-g(x)\}^2}{\lambda^2}\right] \\
&= a^2 A - 2aB + C. \\
\text{Therefore, } p(a|x,y) &\propto \exp\left[-\frac{A}{2}\left\{a^2 - 2a\frac{B}{A} + \left(\frac{B}{A}\right)^2\right\}\right] \\
&= \exp\left\{-\frac{A}{2}\left(a-\frac{B}{A}\right)^2\right\}
\end{aligned}$$

$$\begin{aligned}
\text{so } (a|x,y) &\sim N\left(\frac{B}{A}, A^{-1}\right) \\
&= N\left[\frac{\frac{x}{\sigma_a^2}+\frac{\theta\{y-g(x)\}}{\lambda^2}}{\frac{1}{\sigma_a^2}+\frac{\theta^2}{\lambda^2}}, \left(\frac{1}{\sigma_a^2}+\frac{\theta^2}{\lambda^2}\right)^{-1}\right].
\end{aligned}$$

The predicted heart age is the mean of the posterior distribution.

$$\begin{aligned} \hat{a} &= E(a|x,y) = \int ap(a|x,y)da \\ &= \frac{\frac{x}{\sigma_a^2} + \frac{\hat{\theta}\{y-\hat{g}(x)\}}{\hat{\lambda}^2}}{\frac{1}{\sigma_a^2} + \frac{\hat{\theta}^2}{\hat{\lambda}^2}} \\ \text{where } \hat{g}(x) &= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - \hat{\theta} x. \end{aligned}$$

Through algebraic manipulation, we can rewrite the predicted heart age as

$$\hat{a} = x + \frac{\frac{\hat{\theta}\left\{y-\hat{\boldsymbol{\beta}}\underline{\boldsymbol{x}}\right\}}{\hat{\lambda}^2}}{\frac{1}{\sigma_a^2} + \frac{\hat{\theta}^2}{\hat{\lambda}^2}},$$

where $\underline{\boldsymbol{x}} = [1, x, x^2]^T$.

Multivariate case To generalize, assume there are k ECG outputs in the model and n subjects. For a given subject, let $\boldsymbol{y} = [y_1, y_2, \ldots, y_k]^T$ and let

$$\begin{aligned} \boldsymbol{\theta} &= [\theta_1, \theta_2, \ldots, \theta_k]^T, \\ \boldsymbol{\beta} &= \begin{bmatrix} \beta_{01} & \beta_{11} & \beta_{21} \\ \beta_{02} & \beta_{12} & \beta_{22} \\ \vdots & \vdots & \vdots \\ \beta_{0k} & \beta_{1k} & \beta_{2k} \end{bmatrix}, \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11}^2 & \lambda_{12}^2 & \ldots & \lambda_{1k}^2 \\ \lambda_{12}^2 & \lambda_{22}^2 & \ldots & \lambda_{2k}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{1k}^2 & \lambda_{2k}^2 & \ldots & \lambda_{kk}^2 \end{bmatrix} \end{aligned}$$

and let $g(\underline{\boldsymbol{x}}, x) = \boldsymbol{\beta}\underline{\boldsymbol{x}} - x\boldsymbol{\theta}$. So, $(\boldsymbol{y}|x,a) = N\left(\boldsymbol{\beta}\underline{\boldsymbol{x}} + (a-x)\boldsymbol{\theta}, \boldsymbol{\Lambda}\right)$ and the posterior

distribution of heart age,

$$\begin{aligned}
p(a|x,\boldsymbol{y}) &\propto p(a|x)p(\boldsymbol{y}|x,a) \\
&= \exp\left\{-\frac{1}{2}Q(a)\right\}, \text{ where} \\
Q(a) &= \frac{1}{\sigma_a^2}\{a-x\}^2 + \{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)-a\boldsymbol{\theta}\}^T\boldsymbol{\Lambda}^{-1}\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)-a\boldsymbol{\theta}\} \\
&= \frac{a^2}{\sigma_a^2} - 2a\frac{x}{\sigma_a^2} + \frac{x^2}{\sigma_a^2} \\
&\quad + [a\boldsymbol{\theta}-\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\}]^T\boldsymbol{\Lambda}^{-1}[a\boldsymbol{\theta}-\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\}] \\
&= \frac{a^2}{\sigma_a^2} - 2a\frac{x}{\sigma_a^2} + \frac{x^2}{\sigma_a^2} + a^2\boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta} - 2a\boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\} \\
&\quad + \{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\}^T\boldsymbol{\Lambda}^{-1}\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\} \\
&= a^2\left[\frac{1}{\sigma_a^2} + \boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}\right] - 2a\left[\frac{x}{\sigma_a^2} + \boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\}\right] + C \\
&= a^2A - 2aB + C. \\
\text{Therefore, } p(a|x,\boldsymbol{y}) &\propto \exp\left\{-\frac{A}{2}\left(a-\frac{B}{A}\right)^2\right\} \\
\text{so } (a|x,\boldsymbol{y}) &\sim N\left(\frac{B}{A}, A^{-1}\right) \\
&= N\left[\frac{\frac{x}{\sigma_a^2} + \boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\{\boldsymbol{y}-g(\underline{\boldsymbol{x}},x)\}}{\frac{1}{\sigma_a^2} + \boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}}, \frac{1}{\frac{1}{\sigma_a^2} + \boldsymbol{\theta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}}\right]
\end{aligned}$$

and the predicted heart age,

$$\hat{a} = \frac{\frac{x}{\sigma_a^2} + \hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\Lambda}}^{-1}\{\boldsymbol{y}-\hat{g}(\underline{\boldsymbol{x}},x)\}}{\frac{1}{\sigma_a^2} + \hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{\theta}}} \text{ where } \hat{g}(\underline{\boldsymbol{x}},x) = \hat{\boldsymbol{\beta}}\underline{\boldsymbol{x}} - \hat{\boldsymbol{\theta}}x.$$

Through algebraic manipulation, we can rewrite this as

$$\hat{a} = x + \frac{\hat{\boldsymbol{\theta}}^T\hat{\boldsymbol{\Lambda}}^{-1}\left\{\boldsymbol{y}-\hat{\boldsymbol{\beta}}\underline{\boldsymbol{x}}\right\}}{\frac{1}{\sigma_a^2} + \hat{\boldsymbol{\theta}}\hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{\theta}}}.$$

To obtain $\hat{\boldsymbol{\beta}}$, let $\boldsymbol{y}_i$ be the vector of k ECG outputs for the ith subject and let x_i be the body age of the ith subject, where $i = 1, 2, \ldots, n$.

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ik} \end{bmatrix}, \underline{\boldsymbol{x}}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{11} & \beta_{21} \\ \beta_{02} & \beta_{12} & \beta_{22} \\ \vdots & \vdots & \vdots \\ \beta_{0k} & \beta_{1k} & \beta_{2k} \end{bmatrix}$$

The multivariate analog of the model for $(y|x)$ is

$$\boldsymbol{y}_i \;\; = \;\; \boldsymbol{\beta}\underline{\boldsymbol{x}}_i + \boldsymbol{\nu}_i$$

where $\text{Var}(\boldsymbol{\nu}_i) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Then, $(\boldsymbol{y}_1|x_1), (\boldsymbol{y}_2|x_2), \ldots, (\boldsymbol{y}_n|x_n) \sim N(\boldsymbol{\beta}\underline{\boldsymbol{x}}_i, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is the k by k covariance matrix of the ECG outputs and $\text{Cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \mathbf{0}$.

Let

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \ldots & y_{1k} \\ y_{21} & y_{22} & \ldots & y_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \ldots & y_{nk} \end{bmatrix}, \boldsymbol{E} = \begin{bmatrix} e_{11} & e_{12} & \ldots & e_{1k} \\ e_{21} & e_{22} & \ldots & e_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \ldots & e_{nk} \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{11}^2 \\ 1 & x_{21} & x_{21}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n1}^2 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1k}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 & \ldots & \sigma_{2k}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1k}^2 & \sigma_{2k}^2 & \ldots & \sigma_{kk}^2 \end{bmatrix}.$$

Then, $\hat{\boldsymbol{\beta}}^T = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$. If $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}$, we estimate $\boldsymbol{\Sigma}$ as follows: $\hat{\sigma}_{rs}^2 = \frac{1}{(n-3)}\sum_{i=1}^{n} e_{ir}e_{is}$. However, we still need estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ to get the predicted

heart age.

$$\begin{aligned} \text{Var}(\boldsymbol{y}_i|x_i) &= \boldsymbol{\Sigma} = \text{Var}\left\{(a_i - x_i)\boldsymbol{\theta} + \boldsymbol{\epsilon}_i\right\} \\ &= \sigma_a^2 \boldsymbol{\theta}\boldsymbol{\theta}^T + \boldsymbol{\Lambda}. \end{aligned}$$

As an approximation, ignoring the off-diagonal elements of $\boldsymbol{\theta}\boldsymbol{\theta}^{\boldsymbol{T}}$, we estimated $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \pm\sqrt{\text{diag}\left\{\frac{1}{\sigma_a^2}\left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Lambda}}\right)\right\}},$$

where the sign of $\hat{\theta}_k$ is the sign of $\hat{\beta}_{1k}$.

We specified $\sigma_a^2 = 7.5$ in the prior of heart age so if we can get an estimate of $\boldsymbol{\Lambda}$, we can get an estimate of $\boldsymbol{\theta}$. To do so, we use the repeated measures data (Batdorf et al., 2006). There are two measurements, taken a month apart, for each of the $m = 15$ subjects. A more robust estimate of $\boldsymbol{\Lambda}$ could be achieved with more subjects in the repeated measures study. Let $d_{ik} = y_{ik1} - y_{ik2}$ be the difference between the measurements for the ith subject and kth ECG output, so that

$$\begin{aligned} d &= \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} \\ d_{21} & d_{22} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ d_{m1} & d_{m2} & \dots & d_{mk} \end{bmatrix} \\ \text{so, } \hat{\lambda}_{rs}^2 &= \frac{1}{2m}\sum_{i=1}^{m} d_{ir}d_{is}, \text{ for } r = 1, 2, \dots, k, s = 1, 2, \dots, k. \text{ Thus,} \\ \hat{\boldsymbol{\theta}} &= \sqrt{\frac{1}{\sigma_a^2}\text{diag}\left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Lambda}}\right)}. \end{aligned}$$

5.4 Results

Experts agree that women's hearts age differently than men's hearts. Women's hearts tend to age slower than men's until post menopause, at which time their cardiovascular age quickly catches up to males. Because of this phenomenon, we need to check the model for a gender effect. To determine the need for a gender-specific model, we performed a t-test on the differences between the predicted heart ages of the gender-specific model and the model that does not take gender into account. If we do not take gender into account, the model is biased upwards for males (predicts higher heart ages) and biased downwards for females (predicts lower heart ages) so we designed a gender-specific model (each gender has its own model). If we let $z = 1$ if the subject is male and $z = 0$ if the subject is female, the posterior for heart age in the gender specific model is:

$$\begin{aligned} p(a|x,y,z) &= \frac{p(a|x)p(y|x,a,z)}{\int p(a|x)p(y|x,a,z)da}, \text{ and} \\ \hat{a} &= z\left[x + \frac{\hat{\boldsymbol{\theta}}_1^T \hat{\boldsymbol{\Lambda}}^{-1}\left\{\boldsymbol{y} - \hat{\boldsymbol{\beta}}_1 \underline{\boldsymbol{x}}\right\}}{\frac{1}{\sigma_a^2} + \hat{\boldsymbol{\theta}}_1^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\boldsymbol{\theta}}_1}\right] + (1-z)\left[x + \frac{\hat{\boldsymbol{\theta}}_2^T \hat{\boldsymbol{\Lambda}}^{-1}\left\{\boldsymbol{y} - \hat{\boldsymbol{\beta}}_2 \underline{\boldsymbol{x}}\right\}}{\frac{1}{\sigma_a^2} + \hat{\boldsymbol{\theta}}_2^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\boldsymbol{\theta}}_2}\right]. \end{aligned}$$

We used two linear combinations of ECG variables, $y_1 = \boldsymbol{\gamma_1}^T \boldsymbol{v}_1$ and $y_2 = \boldsymbol{\gamma_2}^T \boldsymbol{v}_2$ as outcomes so that $\boldsymbol{y} = [y_1, y_2]^T$. The first ECG variable, y_1, varies with age and disease and y_2 varies with disease.

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ \text{Taxis} \\ \text{Pd} \\ \sin\left(\text{FrQRSMax} * \pi/180\right) \\ \text{LnHF} \\ \text{LnRMSsum} \\ \text{LnSpatialJT} \end{bmatrix} \text{ and } \boldsymbol{v}_1 = \begin{bmatrix} 1 \\ \text{IIQTVI} \\ \text{V5UnexQTVI} \\ \sin\left(\text{QRSaxis} * \pi/180\right) \\ \text{Pd} \\ \text{MeanAngle} \\ \text{LnnTV} \end{bmatrix}$$

$$\boldsymbol{\gamma}_1 = \begin{bmatrix} -33.0669357 \\ -0.007471284 \\ 0.0524961 \\ -3.977162174 \\ -0.75406667 \\ 0.295048301 \\ 5.607131563 \end{bmatrix} \text{ and } \boldsymbol{\gamma}_2 = \begin{bmatrix} -5.561987914 \\ 3.278798871 \\ 1.482313958 \\ -2.6315664 \\ 0.090181799 \\ 0.048045487 \\ 1.426993361 \end{bmatrix}$$

The resulting estimates for the gender-specific model (1 corresponds to males, 2

corresponds to females) were:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_1 &= [0.175660, 0.274744]^T, \\
\hat{\boldsymbol{\theta}}_2 &= [0.144594, 0.243787]^T, \\
\hat{\boldsymbol{\beta}}_1 &= \begin{bmatrix} -6.209620 & 0.145773 & -0.000680 \\ -10.826300 & 0.231253 & -0.001750 \end{bmatrix}, \\
\hat{\boldsymbol{\beta}}_2 &= \begin{bmatrix} -4.794610 & 0.066861 & 2.37 \times 10^{-6} \\ -7.84566 & 0.083569 & -0.000127 \end{bmatrix}, \text{ and} \\
\hat{\boldsymbol{\Lambda}}^{-1} &= \begin{bmatrix} 4.649546 & -0.064033 \\ -0.064033 & 0.519129 \end{bmatrix}.
\end{aligned}
$$

While the assumptions of the model are based on the healthy non-athlete populations, it is useful to see how the model performs under other conditions, such as when the subject has risk factors or cardiac disease. The results of the gender-specific model are shown below in Figures 5.2 through 5.6. Normally, when making predictions, we want the predicted values to be equal to the observed values. In the figures below, this is symbolized by heart age equals body age (red line) where the x-axis is the subject's body age and the y-axis is the subject's predicted heart age. If the subject's body age equals the subject's predicted heart age, it will fall on the red line. If the subject's predicted heart age is higher than the subject's body age, it will be above the red line and if the subject's predicted heart age is lower than the subject's body age, it will be below the red line. In this case, we want predicted heart age to be centered around the subject's body age, but not necessarily equal to the subject's body age because we want to take the subject's ECG into account when computing heart age. In the training data, we see this for both females and males (Figure 5.2). Predicted heart ages are centered around the red line with some

variability according to each subject's ECG. We also observe this phenomena in the test set, Figure 5.3.

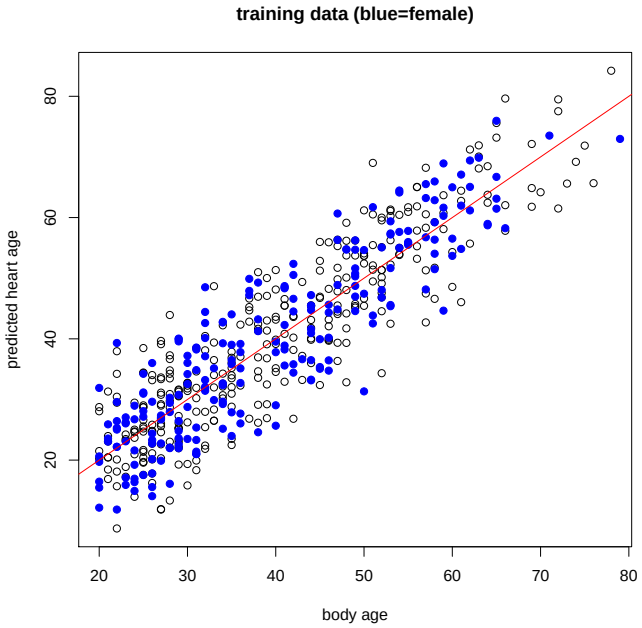

Figure 5.2: Body age versus predicted heart age in the training set. Black circle = male. Blue dot = female.

In subjects with risk factors (Figure 5.4), we expect that most will have higher predicted heart ages than their respective body ages but this should not be the case for everyone. Just because a person has diabetes does not mean that he/she also has heart disease. And, in subjects with heart disease (Figure 5.5), we expect most subjects to have a higher predicted heart age than their body age. Indeed, these results bear this out.

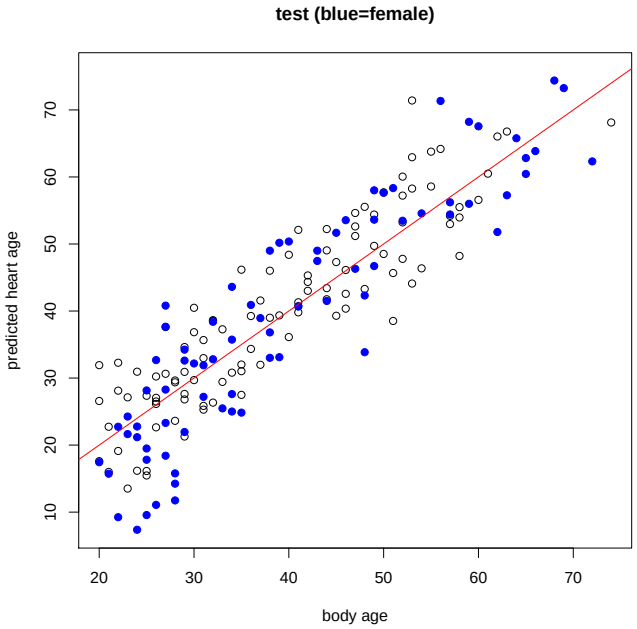

Figure 5.3: Body age versus predicted heart age in the test set. Black circle = male. Blue dot = female.

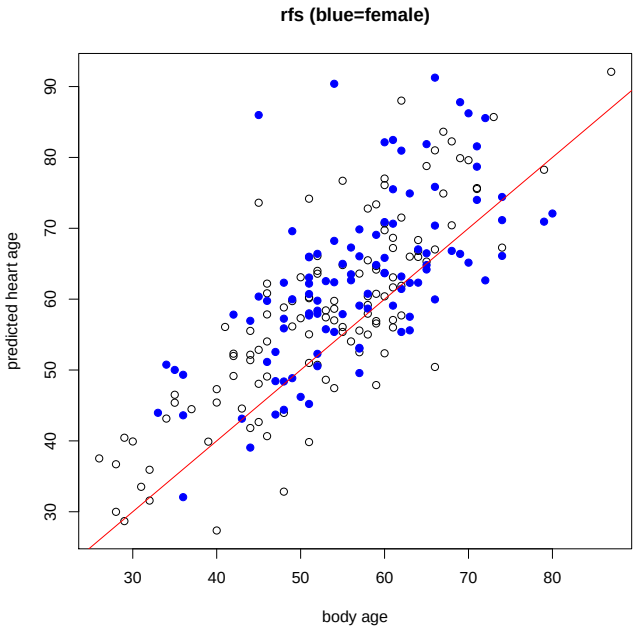

Figure 5.4: Body age versus predicted heart age for subjects with risk factors. Black circle= male. Blue dot = female.

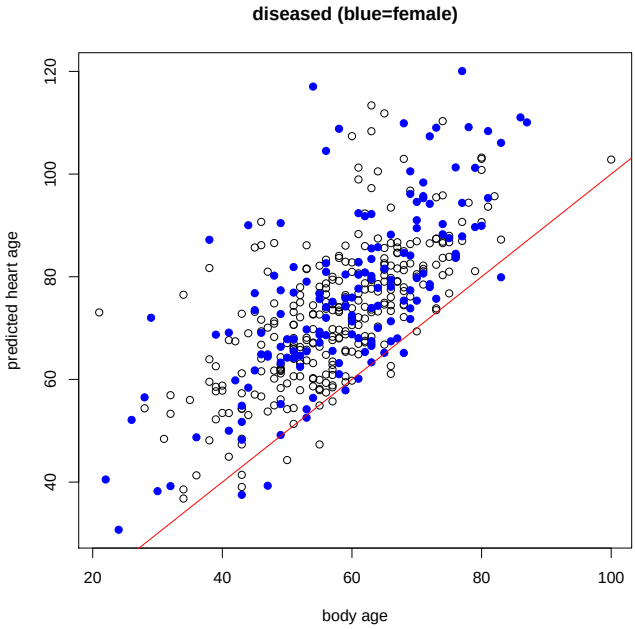

Figure 5.5: Body age versus predicted heart age for subjects with disease. Black circle= male. Blue dot = female.

We also considered the healthy athlete group (Figure 5.6). Interestingly, we did not find that athletes had lower predicted heart ages than their body age. In fact, 56.25% of the athletes had a predicted heart age higher than their body age whereas 51% of subjects in the healthy non-athlete training set had a predicted heart age higher than their body age. Recall that these athletes are endurance-trained elite athletes. A recent Mayo Clinic study (O'Keefe et al., 2012) found that these types of athletes may actually do damage to their hearts by the manner in which they train. Our predictions appear to support this conjecture.

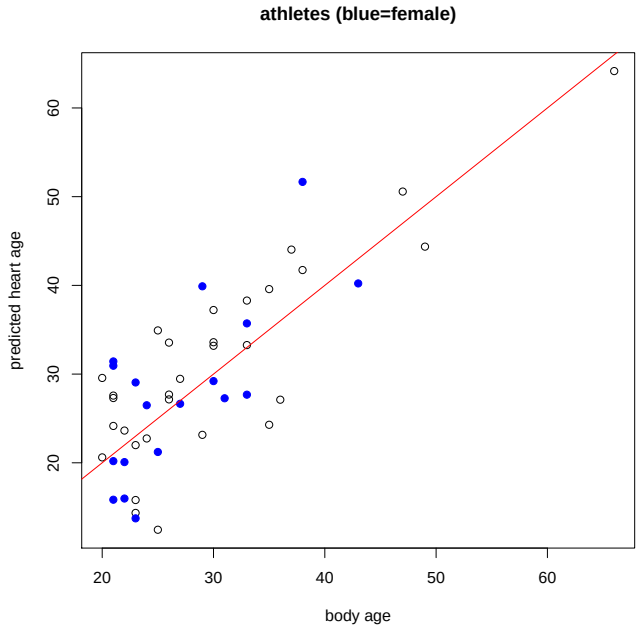

Figure 5.6: Body age versus predicted heart age for athletes. Black circle = male. Blue dot = female.

5.5 Discussion and conclusion

The model was designed for healthy individuals at least 20 years old. However, it is useful to see how the model performs in subjects with risk factors, elite endurance trained athletes, and subjects with heart disease. Heart age predictions in healthy non-athletes are centered around body age but take the variability of each subject's ECG into account. Heart age predictions in other subgroups are consistent with current knowledge. About three-fourths of the subjects with risk factors have higher predicted heart ages than their body ages and almost all of the subjects with disease have higher predicted heart ages than their body ages. Furthermore, the model seems to predict possible anomalies in endurance-trained athletes' ECGs. In our sample, 56.25% of these athletes have predicted heart ages than exceed their body ages.

One possible problem with our method is that we may find that $\hat{\lambda}^2 > \hat{\sigma}_\nu^2$, in which

case, $\hat{\theta}$ is not defined. An alternative approach is to use a Bayesian model for θ with a prior centered on β_1.

While we do not have heart age measurements with which to compare our predictions, Todd Schlegel, a physician at NASA Johnson Space Center and expert in ECG science, confirmed that the heart age predictions are consistent with the current knowledge of ECG measurements and heart health. This model has the potential to aid in the diagnosis of symptomatic patients who undergo an ECG. It could give the physician and patient additional information regarding the patient's cardiac health, and this could lead to further testing or a change in lifestyle. Imagine a patient who has not been diagnosed with heart disease but who has a higher predicted heart age than the patient's body age. This news could motivate the patient to eat healthier and exercise and could also signify the need for further testing, perhaps catching a disease before it gets worse.

6. CONCLUSIONS

This dissertation consists of four main projects in the areas of proteomics, genomics, and cardiology. I have developed a data-based method for protein subcellular localization. It compares well with the current gold-standard procedure, PSORTb v.3.0 (Yu et al., 2010). In addition, the simulation study produced excellent results and it is a flexible method that can be applied to a wide variety of organisms. In the area of genomics, I have completed two projects (Li et al., 2012; Yuan et al., 2012) with scientists at the University of Texas MD Anderson Cancer Center. In the first study, we investigated the predictability of lethality of the knockout mouse using genomic features and in the second study, we analyzed breakpoint hotspots across eight cancer types using the random forest technique. Lastly, I have developed a statistical model that predicts a subject's heart age based on electrocardiogram measurements.

REFERENCES

Ashley, E. A., Raxwal, V., and Froelicher, V. (2001), "An evidence-based review of the resting electrocardiogram as a screening technique for heart disease," *Progress in Cardiovascular Diseases*, 44, 55–67.

Barber, R. D., Rott, M. A., and Donohue, T. J. (1996), "Characterization of a glutathione-dependent formaldehyde dehydrogenase from Rhodobacter sphaeroides," *J. Bacteriology*, 178, 1386–1393.

Batdorf, B. H., Feiveson, A. H., and Schlegel, T. T. (2006), "The effect of signal averaging on the reproducibility and reliability of measures of T-wave morphology," *Journal of Electrocardiology*, 39, 266–270.

Breiman, L. (2001), "Random forests," *Machine Leanring*, 45, 5–32.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Eerdewegh, P. V. (2005), "Identifying SNPs predictive of phenotype using random forests," *Genetic Epidemiology*, 28, 171–182.

Callister, S. J., Dominguez, M. A., Nikora, C. D., Zeng, X., Tavano, C. L., Kaplan, S., Donohue, T. J., Smith, R. D., and Lipton, M. S. (2006a), "Application of the accurate mass and time approach to the proteome analysis of sub-cellular fractions obtained from Rhodobacter sphaeroides 2.4.1. aerobic and photosynthetic cell cultures," *J. Proteome Research*, 5, 1940–1947.

Callister, S. J., Nicora, C. D., Zeng, X. H., Roh, J. H., Dominguez, M. A., Tavano, C. L., Monroe, M. E., Kaplan, S., Donohue, T. J., Smith, R. D., and Lipton, M. S.

(2006b), "Comparison of aerobic and photosynthetic Rhodobactersphaeroides 2.4.1 proteomes," *J. Microbiol. Meth.*, 67, 424–436.

Cohen-Bazire, G., Sistrom, W. R., and Stanier, R. Y. (1957), "Kinetic studies of pigment synthesis by non-purple sulfur bacteria," *J. Cell. Comp. Physiol.*, 49, 25–68.

Collins, F. S., Rossant, J., and Wurst, W. (2007), "A mouse for all reasons," *Cell*, 128, 9–13.

Deal, C. D. and Kaplan, S. (1983), "Immunochemical relationship of the major outer membrane protein of Rhodopseudomonas sphaeroides 2.4.1 to proteins of other photosynthetic bacteria," *J. Bacteriology*, 154, 1015–1020.

Flory, J. E. and Donohue, T. J. (1995), "Organization and expression of the Rhodobacter sphaeroides cycFG operon," *J. Bacteriology*, 177, 4311–4320.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Hsu, J. (2006), *Multiple comparisons: theory and methods*, London: Chapman and Hall.

Jackson, J. B. (1991), "The proton-translocating nicotinamide adenine dinucleotide transhydrogenase," *J. Bioenerg. Biomembr.*, 23, 715–741.

Kardys, I., Kors, J. A., van der Meer, I. M., Hofman, A., van der Kuip, D. A., and Witteman, J. C. (2003), "Spatial QRS-T angle predicts cardiac death in a general population," *European Heart Journal*, 24, 1357–1364.

Levy, D., Labib, S. B., Anderson, K. M., Christiansen, J. C., Kannel, W. B., and Castelli, W. P. (1990), "Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy," *Circulation*, 81, 815–820.

Li, Y., Zhang, L., Ball, R. L., Liang, X., Li, J., Lin, Z., and Liang, H. (2012), "Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots," *Human Molecular Genetics*, 21, 4957–4965.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22.

Monahan, J. F. (2008), *A primer on linear models*, Boca Raton, FL: Chapman and Hall/CRC.

O'Keefe, J. H., Patil, H. R., Lavie, C. J., Magalski, A., Vogel, R. A., and McCullough, P. A. (2012), "Potential Adverse Cardiovascular Effects From Excessive Endurance Exercise," *Mayo Clinic Proceedings*, 87, 587–595.

Poplack Potter, S. L., Holmqvist, F. H., Platonov, P. G., Steding, K., Arheden, H., Pahlm, O., Starc, V., McKenna, W. J., and Schlegel, T. T. (2010), "Detection of hypertrophic cardiomyopathy is improved when using advanced rather than strictly conventional 12-lead electrocardiogram," *Journal of Electrcardiology*, 43, 713–718.

Schlegel, T. T., Kulecz, W. B., DePalma, J. L., Feiveson, A. H., Wilson, J. S., Rahman, M. A., and Bungo, M. W. (2004), "Real-time 12-lead high frequency QRS electrocardiography for enhanced detection of myocardial ischemia and coronary heart disease," *Mayo Clinic Proceedings*, 79, 339–350.

Schlegel, T. T., Kulecz, W. B., Feiveson, A. H., Greco, E. C., DePalma, J. L., Starc, V., Vrtovec, B., Rahman, M. A., Bungo, M. W., Hayat, M. J., Bauch, T., Delgado, R., Warren, S. G., Nunez-Medina, T., Medina, R., Jugo, D., Arheden, H., and Pahlm, O. (2010), "Accuracy of advanced versus strictly conventional 12-lead ECG for detection and screening of coronary artery disease, left ventricular hypertrophy and left ventricular systolic dysfunction," *BMC Cardiovascular Disorders*, 10, 28.

Sox, Jr, H. C., Garber, A. M., and Lettenberg, B. (1989), "The resting electrocardiogram as a screening test. A clinical analysis," *Ann Intern Med*, 111, 489–502.

Stanley, J. R., Adkins, J. N., Slysz, G. W., Monroe, M. E., Purvine, S. O., Karpievitch, Y. V., Anderson, G. A., Smith, R. D., and Dabney, A. R. (2011), "A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics," *Anal Chem.*, 83, 6135–6140.

Tai, S. B. and Kaplan, S. (1985), "Intracellular localization of phospholipid transfer activity in Rhodopseudomonas sphaeroides and a possible role in membrane biogenesis," *J. Bacteriology*, 164, 181–186.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso," *J. Royal Statistical Society. Series Bl*, 58, 267–288.

Weiss, R. L. (1976), "Protoplast formation in Escherichia coli," *J. Bacteriology*, 128, 668–670.

Welch, B. L. (1947), "The generalization of "Student's" problem when several different population variances are involved," *Biometrika*, 34, 28–35.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., and Brinkman, F. S. L. (2010), "PSORTb

3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, 26, 1608–1615.

Yuan, Y., Xu, Y., Xu, J., Ball, R. L., and Liang, H. (2012), "Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data," *Bioinformatics*, 28, 1246–1252.

Zeng, X., Roh, J. H., Callister, S. J., Tavano, C. L., Donohue, T. J., Lipton, M. S., and Kaplan, S. (2007), "Proteomic characterization of the Rhodobacter sphaeroides 2.4.1 photosynthetic membrane: identification of new proteins," *J Bacteriol.*, 189, 7464–7474.